Model-Assisted Uniformly Honest Inference for Optimal Treatment Regimes in High Dimension

Yunan Wu, Lan Wang and Haoda Fu

Abstract

This paper develops new tools to quantify uncertainty in optimal decision making and to gain insight into which variables one should collect information about given the potential cost of measuring a large number of variables. We investigate simultaneous inference to determine if a group of variables is relevant for estimating an optimal decision rule in a high-dimensional semiparametric framework. The unknown link function permits flexible modeling of the interactions between the treatment and the covariates, but leads to nonconvex estimation in high dimension and imposes significant challenges for inference. We first establish that a local restricted strong convexity condition holds with high probability and that any feasible local sparse solution of the estimation problem can achieve the near-oracle estimation error bound.

[1]Yunan Wu is Assistant Professor, Department of Mathematical Sciences, University of Texas at Dallas. Emails: yunan.wu@utdallas.edu. Lan Wang is Professor, Department of Management Science, University of Miami. Emails: lanwang@mbs.miami.edu. Wang and Wu's research was partly supported by NSF DMS-1952373 and NSF OAC-1940160. Dr. Haoda Fu is Research Fellow, Enterprise Lead for Machine Learning and AI, Eli Lilly and Company. Email: fu_haoda@lilly.com. Wang and Wu's research was partly supported by NSF DMS-1952373 and NSF OAC-1940160. The authors are grateful to the referees, the associate editor and the Co-editor for their valuable comments, which have significantly improved the paper.

We further rigorously verify that a wild bootstrap procedure based on a debiased version of the local solution can provide asymptotically honest uniform inference for the effect of a group of variables on optimal decision making. The advantage of honest inference is that it does not require the initial estimator to achieve perfect model selection and does not require the zero and nonzero effects to be well-separated. We also propose an efficient algorithm for estimation. Our simulations suggest satisfactory performance. An example from a diabetes study illustrates the real application.

Keywords: confidence interval; inference; kernel smoothing; multiplier bootstrap; high-dimensional data; optimal treatment regime; precision medicine.

1 Introduction

Precision medicine is an innovative practice for disease treatment that takes into account individual variability in genes, environment, and lifestyle for each patient. Substantial efforts have recently been devoted to studying how to estimate the optimal personalized treatment regime given the individual-level information, which aims to yield the best expected outcome if the treatment regime is followed by each individual in the population. Several successful approaches have been developed for this estimation problem, including Q-learning and A-learning based methods [Watkins and Dayan, 1992, Robins et al., 2000, Murphy, 2003, Moodie and Richardson, 2010, Qian and Murphy, 2011], and classification-based methods [Zhang et al., 2012, Zhao et al., 2012, 2015, Wang et al., 2018, Qi et al., 2018], among others. We refer to Chakraborty and Moodie [2013] and Kosorok and Moodie [2016] for a general introduction to this area and other relevant references.

Inference or uncertainty quantification is important in practice. This paper studies the following inference problem for optimal personalized decision making: suppose we have a large number of covariates (e.g., hundreds of genes), how will we determine if a given subset of covariates (e.g., genes associated with a given biological pathway) is relevant for making the optimal treatment recommendation? Scientifically, this knowledge would enable the doctors and researchers to identify critical characteristics (e.g., gender, age, gene pathways) that are influential for the optimal decision. It also helps gain insight into what information is worth collecting to be more cost effective, given the possibility of measuring a large number of variables (genetic, clinic, etc).

In the last few years, important progress has been made in inference with optimal decision rules. Laber et al. [2014] developed a novel locally consistent adaptive confidence interval for the Q-learning approach. Chakraborty et al. [2013] proposed a practically convenient adaptive m-out-of-n bootstrap method for inference for Q-learning. Song et al. [2015] studied penalized Q-learning. Jeng et al. [2018] developed Lasso-based debiased procedure for A-learning. Different but related, Chakraborty et al. [2014] and Luedtke and van der Laan [2016], Zhu et al. [2019] developed confidence intervals for another quantity of interest: the value function. However, existing work mostly deals with the classical asymptotic setting of fixed p and large n, where p is the number of covariates and n is the sample size, and have not addressed the challenge of inference with high-dimensional variables. Moreover, the aforementioned work often assumes that the interaction between the covariates and the treatment has a known functional form.

Motivated by the overarching goal of precision medicine to incorporate genetic information (e.g, measurements on thousands of genes) in the decision making process, this paper investigates inference about the effect of a group of variables on the optimal decision rule in the high-dimensional setting. The existing frameworks are known to face challenges for the purpose of inference in high dimension. The Q-learning approach is prone to model-misspecification. Robust model-free procedures that directly estimate the Bayes rule (e.g., Zhang et al. [2012]) have a nonstandard convergence rate, see for example, the recent analysis in Wang et al. [2018] on the cubic-root convergence rate. On the other hand, the theory of Hinge-loss based O-learning (Zhao et al. [2012]) has been focused on the generalization error bound. Inference for the Bayes rule based on the nonsmooth surrogate loss is very challenging in high dimension. We alleviate the above difficulty by adopting a flexible semiparametric model-assisted approach for optimal decision estimation and inference. The semiparametric structure permits nonparametric main effects and nonlinear interaction effect between the covariates and treatment via an unknown smooth link function. This semiparametric framework incorporates many existing models as special cases.

When the interaction effects are nonlinear, the parameter indexing the optimal decision rule does not necessarily correspond to the solution of a convex problem. For inference, we first propose and study a preliminary estimator based on a high-dimensional penalized profile estimation equation. This estimator is motivated by earlier work on classical single-index models (e.g., Powell et al. [1989], Duan and Li [1991], Ichimura [1993], Zhu and Xue [2006], Carroll et al. [1997], Xia et al. [1999], Yu and Ruppert [2002], Wang et al. [2010], Ma and Zhu [2013], Ma and He [2016], among others). Several paper recently studied estimation for high-dimensional single-index models (e.g., Radchenko [2015], Neykov et al. [2016], Yang et al. [2017], Lin et al. [2019], among others) but focused on statistical properties of the global solution which may not be numerically achieved due to the nonconvex nature of the problem. Adopting tools from modern empirical process and random matrix theory, we establish that a local restricted strong convexity condition holds with high probability in high dimension and that any local sparse solution of the penalized estimation equation can achieve desirable estimation accuracy. Moreover, we propose a new algorithm for efficient computation in high dimension.

Our research also makes new contributions to statistical inference in high-dimensional semiparametric models. Recent work on inference has been mostly limited to linear regression or generalized linear regression, see Zhang and Zhang [2014], Van de Geer et al. [2014], Javanmard and Montanari [2014], Belloni et al. [2015], Cai et al. [2017], Ning et al. [2017], Zhang and Cheng [2017], Zhu and Bradic [2018], Shi et al. [2020], among others. High-dimensional inference in the semiparametric setting with estimated nonparametric components is a substantially harder problem and has been little studied. We have a particularly challenging setting where the parameter of interest and nonparametric component are bundled together, that is, the nuisance functions depend on the parameter of interest (Ding and Nan [2011]). So far, statistical inference for single-index model has mostly been limited to the lower-dimensional setting (e.g., Liang et al. [2010]), Gueuning and Claeskens [2016]).

Our approach is inspired by the de-biasing (or de-sparsifying) idea proposed in Zhang and Zhang [2014] and Van de Geer et al. [2014], which intuitively can be thought of inverting the Karush-Kuhn-Tucker conditions [Van de Geer et al., 2014]. We generalize this idea to the semiparametric setting and prove that valid honest uniform inference can be obtained based on a debiased version of a local solution. Specifically, we derive simultaneous confidence intervals for inference on a group of variables while allowing the number of covariates to exceed the sample size. The confidence intervals enjoy the *honest* property in the following sense

$$\sup_{\boldsymbol{\beta}_0:||\boldsymbol{\beta}_0||_0\le s}\ \sup_{\alpha\in(0,1)}\left|P\Big(\sqrt{n}\max_{j\in\mathcal{G}}|\widetilde{\beta}_j-\beta_{0j}|\le c^*_{1-\alpha}\Big)-(1-\alpha)\right|=o(1),$$

where $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p})^T$ is the population parameter indexing the optimal treatment regime, $\widetilde{\beta}_j$'s denote debiased estimators that will be introduced later, $\mathcal{G}$ denotes the group of variables of interest, $||\cdot||_0$ denotes the l_0 norm of a vector, and s is a positive integer denoting the sparsity size. The significance of the honest property is that the coverage probability is asymptotically valid uniformly over a class of s-sparse models. An immediate implication is that it relaxes the assumption on signal strength and does not require the zero and nonzero effects to be well-separated (so-called β_{min} condition). In particular, this procedure does not require the initial estimator to achieve perfect model selection. It avoids the problems associated with the nonuniformity of the limiting theory for penalized estimators, see discussions in Li [1989], Pötscher [2009], Van de Geer et al. [2014], McKeague and Qian [2015], among others. It is also worth noting that the number of variables in $\mathcal{G}$ can be either small or large. For example, one may be interested in assessing how a group of genes corresponding to a particular biological pathway, the size of which can be comparable with or even larger than the sample size, affect optimal decision making. The critical value $c^*_{1-\alpha}$ is obtained using a wild bootstrap procedure, which automatically accounts for the dependence of the coordinates for testing component-wise hypotheses and leads to more accurate finite-sample performance.

The remainder of the paper is organized as follows. Section 2 introduces the new methodology. Section 3 studies the statistical properties. Section 4 provides the details on computation and reports numerical results from Monte Carlo studies. Section 5 illustrates the new methods on a real data example from a diabetes study. Section 6 discusses some extensions. The regularity conditions, all the proofs and additional numerical examples are given in the online supplementary material.

2 Methodology

2.1 A Semiparametric Framework

For notational simplicity, we will focus on the binary decision setting. Let $A \in \mathcal{A} = \{0, 1\}$ denote a binary treatment and $\boldsymbol{x} \in \mathcal{X}$ denote a p-dimensional vector of baseline covariates. Let Y denote the outcome of interest. Without loss of generality, we assume a larger value of the outcome is preferred. The observed data consist of $\{(\boldsymbol{x}_i, A_i, Y_i) : i = 1, \cdots, n\}$. We are interested in the setting where $p \gg n$.

A treatment regime is an individualized decision rule that can be represented as a function $d(\boldsymbol{x}) : \mathcal{X} \to \mathcal{A}$. The optimal treatment regime is defined as the decision rule which, if followed by the whole population, will achieve the largest average outcome. Formally, it is defined using the potential outcome framework in causal inference [Neyman, 1990, Rubin, 1978]. Let $Y^*(a)$ be the potential outcome had the subject been assigned to treatment $a \in \{0, 1\}$. Given a treatment regime $d(\boldsymbol{x})$, the corresponding potential outcome is $Y^*(d) = Y^*(1)d(\boldsymbol{x}) + Y^*(0)(1 - d(\boldsymbol{x}))$. The optimal treatment regime is defined as $d^{\text{opt}}(\boldsymbol{x}) = \arg\max_d \mathrm{E}\{Y^*(d)\}$. It is now well known that $d^{\text{opt}}(\boldsymbol{x}) = \arg\max_{a \in \mathcal{A}} \mathrm{E}(Y|\boldsymbol{x}, A = a)$ [Qian and Murphy, 2011].

This paper considers a flexible semiparametric framework for optimal treatment regime estimation and inference in the high-dimensional setting. Specifically, we assume

$$Y_i = g(\boldsymbol{x}_i) + (A_i - 1/2) f_0(\boldsymbol{x}_i^T \boldsymbol{\beta}_0) + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{\beta}_0 = (\beta_{01}, \beta_{02}, \cdots, \beta_{0p})^T$, $g(\boldsymbol{x}_i)$ is the unknown main effect, and $f_0(\cdot)$ is an unknown function that describes the interaction between the treatment and covariates, and the random error ϵ_i satisfies $\mathrm{E}(\epsilon_i|\boldsymbol{x}_i) = 0$, $i = 1, \ldots, n$. For identification purpose, we assume that there exists a relevant covariate which has a continuous density given the other covariates [Ichimura, 1993]. Such an identification condition is required even in the lower-dimensional setting when the true model is known. Without loss of generality, we assume that the first covariate x_1 satisfies this condition and normalize its coefficient β_{01} such that $\beta_{01} = 1$, see Remark (c) in Section S2 of the online supplementary material for more discussions on the identifiability condition. We denote $\mathbb{B}_0 = \{\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T : \beta_1 = 1\}$ as the candidate set for $\boldsymbol{\beta}_0$. Under model (1), the optimal treatment regime is $d^{\mathrm{opt}}(\boldsymbol{x}) = \mathrm{I}\big(f_0(\boldsymbol{x}_i^T\boldsymbol{\beta}_0) > 0\big)$, where $\mathrm{I}(\cdot)$ denotes the indicator function. Note that the class of index rules are popular in practice due to its interpretability.

Existing work on inference for optimal treatment regime is mostly based on a parametric generative model, which is prone to model misspecification. The semiparametric structure alleviates this difficulty. In particular, it allows for possible nonlinear interaction effects between the covariates and treatment. It also circumvents the curse of dimensionality associated with a fully nonparametric model.

Our goal is to estimate $\boldsymbol{\beta}_0$ and make inference on its components in the high-dimensional setting. In the special case $f_0(u) = u$, which is popularly used in practice, the problem can be formulated as a high-dimensional convex estimation problem. However, when f_0 is nonlinear, it generally leads to a high-dimensional nonconvex problem. Both estimation and inference need to overcome new challenges.

2.2 Profiled Semiparametric Estimation

We start with introducing a penalized profiled semiparametric estimation equation for estimating the parameter indexing the optimal treatment regime. We consider data from a random experiment, that is, $P(A_i = 0) = P(A_i = 1) = 1/2$, $i = 1, \dots, n$. Extension to data from observational studies is discussed in Section 6. Inspired by an observation made for the linear model [Tian et al., 2014], we observe

$$2(2A_i - 1)Y_i = f_0(\boldsymbol{x}_i^T\boldsymbol{\beta}_0) + 2(2A_i - 1)\big[\epsilon_i + g(\boldsymbol{x}_i)\big]. \tag{2}$$

Let $\widetilde{Y}_i = 2(2A_i - 1)Y_i$ be the modified response, and let $\widetilde{\epsilon}_i = 2(2A_i - 1)\big[\epsilon_i + g(\boldsymbol{x}_i)\big]$ be the modified error. We have

$$\mathrm{E}\{\widetilde{Y}_i|\boldsymbol{x}_i\} = f_0(\boldsymbol{x}_i^T\boldsymbol{\beta}_0). \tag{3}$$

In the ideal situation where the link function f_0 is known, we have $\boldsymbol{\beta}_0 = \arg\min_{\boldsymbol{\beta}} \mathrm{E}\big[\widetilde{Y}_i - f_0(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^2$. It is noteworthy that for a nonlinear function f_0, the objective function is usually nonconvex in $\boldsymbol{\beta}$. Ichimura [1993] carefully studied the properties of the global minimizer for a semiparametric nonlinear least-squares approach in the classical finite-dimensional setting.

To estimate $\boldsymbol{\beta}_0$ in the high-dimensional setting with an known f_0, we consider a penalized profiled semiparametric estimation equation. In the ideal situation where f_0 is known a prior, $\boldsymbol{\beta}_0$ satisfies the following unbiased estimating equation

$$\mathrm{E}\big\{\big[\widetilde{Y}_i - f_0(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]f_0'(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\boldsymbol{x}_i\big\} = \mathbf{0}, \tag{4}$$

where $f_0'(\cdot)$ denotes the derivative of $f_0(\cdot)$. We will replace the unknown f_0 and f_0' by their respective profiled nonparametric estimator, and consider an appropriately penalized version of the estimated score function to handle the high-dimensional covariates.

We summarize the main steps of estimation as follows. Define $G(t|\boldsymbol{\beta}) = \mathrm{E}\{\widetilde{Y}|\boldsymbol{x}^T\boldsymbol{\beta} = t\}$. Note that $G(t|\boldsymbol{\beta}_0) = f_0(t)$. However, when $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$, $G(t|\boldsymbol{\beta})$ usually has a functional form different from f_0. Ichimura [1993] showed that $\frac{\partial G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \approx f_0'(\boldsymbol{x}_i^T\boldsymbol{\beta})\big[\boldsymbol{x}_i - \mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})\big]^T$ for $\boldsymbol{\beta}$ close to $\boldsymbol{\beta}_0$. Consider the Nadaraya-Watson kernel estimator for $G(t|\boldsymbol{\beta})$:

$$\widehat{G}(t|\boldsymbol{\beta}) = \sum_{i=1}^{n} W_{ni}(t,\boldsymbol{\beta})\widetilde{Y}_i, \tag{5}$$

where $K_h(z) = h^{-1}K(z/h)$, and $W_{ni}(t,\boldsymbol{\beta}) = \frac{K_h(t-\boldsymbol{x}_i^T\boldsymbol{\beta})}{\sum_{j=1}^n K_h(t-\boldsymbol{x}_j^T\boldsymbol{\beta})}$. Write $G^{(1)}(t|\boldsymbol{\beta}) = \frac{d}{dt}G(t|\boldsymbol{\beta})$ and $W_{ni}^{(1)}(t,\boldsymbol{\beta}) = \frac{d}{dt}W_{ni}(t,\boldsymbol{\beta})$. Then the kernel estimator for the derivative $G^{(1)}(t|\boldsymbol{\beta})$ is

$$\widehat{G}^{(1)}(t|\boldsymbol{\beta}) = \sum_{i=1}^{n} W_{ni}^{(1)}(t,\boldsymbol{\beta})\widetilde{Y}_i. \tag{6}$$

Write $G(\boldsymbol{x}^T\boldsymbol{\beta}|\boldsymbol{\beta}) = \mathrm{E}\{\widetilde{Y}|\boldsymbol{x}^T\boldsymbol{\beta}\}$. To estimate $\widehat{G}(\boldsymbol{x}_j^T\boldsymbol{\beta}|\boldsymbol{\beta})$ and $\widehat{G}^{(1)}(\boldsymbol{x}_j^T\boldsymbol{\beta}|\boldsymbol{\beta})$, we employ the following leave-one-out estimators

$$\widehat{G}(\boldsymbol{x}_j^T\boldsymbol{\beta}|\boldsymbol{\beta}) = \sum_{i=1,i\neq j}^{n} W_{nij}(\boldsymbol{x}_j^T\boldsymbol{\beta},\boldsymbol{\beta})\widetilde{Y}_i, \quad \widehat{G}^{(1)}(\boldsymbol{x}_j^T\boldsymbol{\beta}|\boldsymbol{\beta}) = \sum_{i=1,i\neq j}^{n} W_{nij}^{(1)}(\boldsymbol{x}_j^T\boldsymbol{\beta},\boldsymbol{\beta})\widetilde{Y}_i, \tag{7}$$

where $W_{nij}(\boldsymbol{x}_j^T\boldsymbol{\beta},\boldsymbol{\beta}) = \frac{K_h(\boldsymbol{x}_j^T\boldsymbol{\beta}-\boldsymbol{x}_i^T\boldsymbol{\beta})}{\sum_{k\neq j} K_h(\boldsymbol{x}_j^T\boldsymbol{\beta}-\boldsymbol{x}_k^T\boldsymbol{\beta})}$, and $W_{nij}^{(1)}(\boldsymbol{x}_j^T\boldsymbol{\beta},\boldsymbol{\beta}) = \frac{d}{dt}W_{nij}(t,\boldsymbol{\beta})\Big|_{t=\boldsymbol{x}_j^T\boldsymbol{\beta}}$. Similarly, we estimate $\mathrm{E}(\boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\beta}_0)$ by $\widehat{\mathrm{E}}(\boldsymbol{x}_j|\boldsymbol{x}_j^T\boldsymbol{\beta}) = \sum_{i=1,i\neq j}^n W_{ni}(\boldsymbol{x}_j^T\boldsymbol{\beta},\boldsymbol{\beta})\boldsymbol{x}_i$. Denote $\boldsymbol{x}_i = (x_{i,1}, \boldsymbol{x}_{i,-1}^T)^T$. Motivated by the semiparaemtric efficient score derived in Liang et al. [2010], we consider the following profiled semiparametric estimating function

$$\boldsymbol{S}_n(\boldsymbol{\beta},\widehat{G},\widehat{\mathrm{E}}) = -n^{-1}\sum_{i=1}^{n}\big[\widetilde{Y}_i - \widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})[\boldsymbol{x}_{i,-1} - \widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})]. \tag{8}$$

In the high-dimensional setting, the estimating equation $\boldsymbol{S}_n(\boldsymbol{\beta}) = \mathbf{0}$ is ill-posed when $p \gg n$. Let $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)^T = (1, \widehat{\boldsymbol{\beta}}_{-1}^T)^T$ be a solution in $\mathbb{B}_0$ that solves the following penalized semiparametric profiled estimating equation

$$\boldsymbol{S}_n(\boldsymbol{\beta}, \widehat{G}, \widehat{\mathrm{E}}) + \lambda \boldsymbol{\kappa} = \mathbf{0}, \tag{9}$$

where $\lambda > 0$ is a tuning parameter, $\boldsymbol{\kappa} = (\kappa_2, \ldots, \kappa_p)^T \in \partial ||\boldsymbol{\beta}_{-1}||_1$ with $||\boldsymbol{\beta}_{-1}||_1$ denoting the l_1 norm of $\boldsymbol{\beta}_{-1} = (\beta_2, \ldots, \beta_p)^T$ and $\partial ||\boldsymbol{\beta}_{-1}||_1$ denoting the subdifferential of $||\boldsymbol{\beta}_{-1}||_1$, that is $\kappa_j = \mathrm{sign}(\beta_j)$ if $\beta_j \neq 0$, and $\kappa_j \in [-1, 1]$ otherwise, $j = 2, \ldots, p$. In (9), $\widehat{G}$ and $\widehat{\mathrm{E}}$ are evaluated at the corresponding $\boldsymbol{\beta}$ in the estimating equations, hence here they stand for $\widehat{G}(\boldsymbol{x}_i^T \boldsymbol{\beta} | \boldsymbol{\beta})$ and $\widehat{\mathrm{E}}(\boldsymbol{x}_i | \boldsymbol{x}_i^T \boldsymbol{\beta})$, respectively. Note that (9) may have multiple solutions. The theory we develop in Section 3.1 provides a near-optimal error bound for any sparse local solution of the estimating equation. The satisfactory performance of the proposed profiled estimator is demonstrated in the numerical simulations in Section 4.2.

2.3 Inference on the Optimal Decision Rule

To quantify the importance of the covariates on optimal decision making, we will construct confidence intervals for the individual components of $\boldsymbol{\beta}_0 = (1, \boldsymbol{\beta}_{0,-1}^T)^T$ via debiasing a local solution to the semiparametric estimating equation (9). This generalizes the work of debiased confidence intervals for high-dimensional linear regression in Zhang and Zhang [2014] and Van de Geer et al. [2014] to the semiparametric setting where the initial estimator is an estimating equation solution and an estimated infinite-dimensional functional is present. The theory for semiparametric inference in high dimension is highly nontrivial and is carefully studied in Section 3. We further investigate a wild bootstrap procedure for testing a general group hypothesis, which aims to achieve accurate finite-sample performance.

Let $\widehat{\boldsymbol{\beta}} = (1, \widehat{\boldsymbol{\beta}}_{-1}^T)^T$ denote a solution satisfying (9). In the high-dimensional linear regression setting, the main idea of debiased estimator is to invert the Karush–Kuhn–Tucker (KKT) condition of the lasso. Inspired by this idea, we consider the following debiased estimator of $\boldsymbol{\beta}_{0,-1}$:

$$\widetilde{\boldsymbol{\beta}}_{-1} = \widehat{\boldsymbol{\beta}}_{-1} - \widehat{\boldsymbol{\Theta}}^T \boldsymbol{S}_n(\widehat{\boldsymbol{\beta}}, \widehat{G}, \widehat{\mathrm{E}}), \tag{10}$$

where the $(p-1) \times (p-1)$ matrix $\widehat{\boldsymbol{\Theta}}$ is an approximation to the inverse of $\nabla \boldsymbol{S}_n(\widehat{\boldsymbol{\beta}}, \widehat{G}, \widehat{\mathrm{E}})$, the derivative matrix of $\boldsymbol{S}_n(\boldsymbol{\beta}, \widehat{G}, \widehat{\mathrm{E}})$ with respect to $\boldsymbol{\beta}_{-1}$ evaluated at $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$. To construct the approximate inverse $\widehat{\boldsymbol{\Theta}}$, we propose a nodewise Dantzig estimator. Specifically, given the initial estimator $\widehat{\boldsymbol{\beta}}$ and a positive number η, for $j = 2, \cdots, p$, define

$$\boldsymbol{d}_j(\widehat{\boldsymbol{\beta}}, \eta) = \mathop{\arg\min}_{\boldsymbol{v} \in \mathbb{R}^{p-2}} ||\boldsymbol{v}||_1 \text{ s.t. } \left|\left| n^{-1} \sum_{i=1}^n \left[\widehat{G}^{(1)}(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}} | \widehat{\boldsymbol{\beta}})\right]^2 (\widehat{x}_{i,j} - \widehat{\boldsymbol{x}}_{i,-j*}^T \boldsymbol{v}) \widehat{\boldsymbol{x}}_{i,-j*} \right|\right|_\infty \leq \eta, \tag{11}$$

where $||\cdot||_\infty$ denotes the infinity norm of a vector, $\widehat{\boldsymbol{x}}_i = \boldsymbol{x}_i - \widehat{\mathrm{E}}(\boldsymbol{x}_i | \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}})$, $\widehat{x}_{i,j}$ denotes the j^{th} entry of the vector $\widehat{\boldsymbol{x}}_i$, $\widehat{\boldsymbol{x}}_{i,-1}$ denotes the $(p-1)$-subvector of $\widehat{\boldsymbol{x}}_i$ that excludes the 1^{st} entry, and the $\widehat{\boldsymbol{x}}_{i,-j*}$ denotes the $(p-2)$-subvector of $\widehat{\boldsymbol{x}}_i$ that excludes the 1^{st} and j^{th} entries. Furthermore, for $j = 2, \cdots, p$, we define

$$\boldsymbol{\phi}_j(\widehat{\boldsymbol{\beta}}, \eta) = \Big(- \big(\boldsymbol{d}_j(\widehat{\boldsymbol{\beta}}, \eta)\big)^T_{1:(j-2)}, 1, -\big(\boldsymbol{d}_j(\widehat{\boldsymbol{\beta}}, \eta)\big)^T_{(j-1):(p-2)} \Big)^T, \tag{12}$$

$$\tau_j^2(\widehat{\boldsymbol{\beta}}, \eta) = n^{-1} \sum_{i=1}^n \left[\widehat{G}^{(1)}(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}} | \widehat{\boldsymbol{\beta}})\right]^2 \widehat{x}_{i,j} \widehat{\boldsymbol{x}}_{i,-1}^T \boldsymbol{\phi}_j(\widehat{\boldsymbol{\beta}}, \eta), \tag{13}$$

$$\boldsymbol{\theta}_j(\widehat{\boldsymbol{\beta}}, \eta) = \tau_j^{-2}(\widehat{\boldsymbol{\beta}}, \eta) \boldsymbol{\phi}_j(\widehat{\boldsymbol{\beta}}, \eta), \tag{14}$$

where for a vector $\boldsymbol{u} = (u_1, \cdots, u_p)^T$, given $1 \le i \le j \le p$, $(\boldsymbol{u})_{i:j}$ returns the subvector $(u_i, \cdots, u_j)^T$, and for any $i > j$, $(\boldsymbol{u})_{i:j}$ returns the empty vector. For notational simplicity, denote $\widehat{\boldsymbol{d}}_j = \boldsymbol{d}_j(\widehat{\boldsymbol{\beta}}, \eta)$, $\widehat{\tau}_j^2 = \tau_j^2(\widehat{\boldsymbol{\beta}}, \eta)$, and $\widehat{\boldsymbol{\theta}}_j = \boldsymbol{\theta}_j(\widehat{\boldsymbol{\beta}}, \eta)$. The approximate inverse of $\nabla \boldsymbol{S}_n(\widehat{\boldsymbol{\beta}}, \widehat{G}, \widehat{\mathrm{E}})$ is then constructed as

$$\widehat{\boldsymbol{\Theta}} = (\widehat{\boldsymbol{\theta}}_2, ..., \widehat{\boldsymbol{\theta}}_p).$$

The validity of $\widehat{\boldsymbol{\Theta}}$ as an approximation to the inverse of $\nabla \boldsymbol{S}_n(\widehat{\boldsymbol{\beta}}, \widehat{G}, \widehat{\mathrm{E}})$ is given in Lemma 2 of Section 3.2. Section 3 will also present the statistical properties of the debiased estimator $\widetilde{\boldsymbol{\beta}}_{-1} = (\widetilde{\beta}_2, \cdots, \widetilde{\beta}_p)^T$. This then leads to the following asymptotic $100(1-\alpha)\%$ confidence interval for β_{0j},

$$\left\{ \widetilde{\beta}_j - \Phi^{-1}(1-\alpha/2)\big(\widehat{\Sigma}_{jj}/n\big)^{1/2}, \widetilde{\beta}_j + \Phi^{-1}(1-\alpha/2)\big(\widehat{\Sigma}_{jj}/n\big)^{1/2} \right\}, \tag{15}$$

where $j = 2, \cdots, p$, $\Phi^{-1}(\cdot)$ is the quantile function of the standard normal distribution, and $\widehat{\Sigma}_{jj}$ denotes the $(j-1)^{th}$ diagonal entry of $\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}})$, with

$$\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}) \triangleq \widehat{\boldsymbol{\Theta}}^T \Big\{ \frac{1}{n} \sum_{i=1}^n \big[\widetilde{Y}_i - \widehat{G}(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}} | \widehat{\boldsymbol{\beta}})\big]^2 [\widehat{G}^{(1)}(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}} | \widehat{\boldsymbol{\beta}})]^2 \widehat{\boldsymbol{x}}_{i,-1} \widehat{\boldsymbol{x}}_{i,-1}^T \Big\} \widehat{\boldsymbol{\Theta}}. \tag{16}$$

Corollary 1 in Section 3 justifies the asymptotic uniform validity of this marginal confidence interval.

Next, we consider the following more general simultaneous testing problem

$$H_{0,\mathcal{G}} : \beta_{0j} = 0 \text{ for all } j \in \mathcal{G} \quad \text{versus} \quad H_{1,\mathcal{G}} : \beta_{0j} \neq 0 \text{ for some } j \in \mathcal{G}, \tag{17}$$

where $\mathcal{G}$ is a prespecified subset of $\{2,\ldots,p\}$. The size of $\mathcal{G}$ may depend on the sample size n. Such a hypothesis naturally arises in the high-dimensional setting. For example, researchers may want to test whether a gene pathway, consisting of multiple genes for the same biological functions, is important for optimal treatment regime recommendation. For this purpose, we propose an effective bootstrap procedure. Although the asymptotic normal distribution of the debiased estimator (see Theorem 2) allows for construction of confidence intervals for individual coefficients (or fixed-dimensional subvector of coefficients), applying it to make inference for groups of variables when the group size diverges (potentially larger than n) is not straightforward. Moreover, confidence intervals based on the asymptotic distribution have been observed to sometimes lead to undercoverage for nonzero coefficients in finite samples. The bootstrap procedure we study automatically accounts for the dependence structure of the variables in the group and provides more accurate critical value.

When deriving the asymptotic property of the debiased estimator (in the proof of Theorem 2), it is observed that the asymptotic property of $\sqrt{n}(\widetilde{\boldsymbol{\beta}}_{-1}-\boldsymbol{\beta}_{0,-1})$ is determined by the leading term $\sqrt{n}\widehat{\boldsymbol{\Theta}}^T\boldsymbol{S}_n(\boldsymbol{\beta}_0,G,\mathrm{E})$. This suggests that we approximate the distribution of $\sqrt{n}(\widetilde{\beta}_j-\beta_{0j})$, $j=2,\ldots,p$, by the distribution of the following multiplier bootstrap statistic

$$\delta_j^* \quad \triangleq \tfrac{1}{n}\textstyle\sum_{i=1}^n r_i\big[\widetilde{Y}_i-\widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\big]\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{x}}_{i,-1}^T\widehat{\boldsymbol{\theta}}_j, \tag{18}$$

where $r_1,\cdots,r_n$ are i.i.d. standard normal random variables, independent of the data. Let $c_{1-\alpha}^*$ be the upper α-quantile of the distribution of $\max_{j\in\mathcal{G}}|\delta_j^*|$ conditional on the data, which can be easily simulated by generating multiple independent copies of the random weights. We reject the null hypothesis at level α if $\max_{j\in\mathcal{G}}|\widetilde{\beta}_j|>c_{1-\alpha}^*$. The asymptotic validity of the bootstrap procedure is formally established in Section 3. Its performance is demonstrated in the numerical simulations in Section 4.2.

3 Statistical Properties

3.1 Theory for Estimation

Making inference about the optimal treatment regime requires an adequate initial estimator for $\boldsymbol{\beta}_0$. To obtain such an initial estimator in the high-dimensional semiparametric framework, a significant challenge is that the corresponding estimation problem is not necessarily convex. To tackle this, we first establish in Lemma 1 below that the estimated $(p-1)$-dimensional gradient $\boldsymbol{S}_n(\cdot, \widehat{G}, \widehat{\mathrm{E}})$ in (8) possesses an important local restricted strong convexity property with high probability. Theorem 1 then shows that all local sparse solutions within a small neighborhood of $\boldsymbol{\beta}_0$ enjoy a near-optimal error rate under mild conditions. In the sequel, we use $a \vee b$ to denote $\max(a, b)$, and $a \wedge b$ to denote $\min(a, b)$. Let $s = ||\boldsymbol{\beta}_0||_0$ be the sparsity size of $\boldsymbol{\beta}_0$, the population parameter indexing the optimal treatment regime.

Lemma 1. *(local restricted strong convexity property) Assume conditions (A1)–(A5) in Section S2 of the online supplementary material are satisfied. If* $d_0\left[\frac{s\log(p\vee n)}{n}\right]^{1/5} \leq h < 1$ *for some constant* $d_0 > 0$*, then there exist universal positive constants* c_0*,* c_1*,* c_2 *and* $r \leq 1$*, which do not depend on* n*,* p *and* $\boldsymbol{\beta}_0$*, such that*

$$\begin{aligned} & P\Big(\big\langle \boldsymbol{S}_n(\boldsymbol{\beta}, \widehat{G}, \widehat{\mathrm{E}}) - \boldsymbol{S}_n(\boldsymbol{\beta}_0, \widehat{G}, \widehat{\mathrm{E}}), \boldsymbol{\beta}_{-1} - \boldsymbol{\beta}_{0,-1}\big\rangle \geq c_0||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_2^2 - c_1 h^2||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_2,\ \forall\ \boldsymbol{\beta} \in \mathbb{B}\Big) \\ \geq\ & 1 - \exp(-c_2 \log p), \end{aligned}$$

for all n *sufficiently large, where* $\mathbb{B} = \{\boldsymbol{\beta} \in \mathbb{B}_0 : ||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_2 \leq r, ||\boldsymbol{\beta}||_0 \leq ks\}$ *and* $k > 1$ *is a* *positive constant.*

Remark 1. Lemma 1 characterizes the local geometry of the profiled score function. For high-dimensional regression with convex loss function such as L_1 penalized linear regression, restricted strong convexity plays an important role on the theory of the regularized estimator Negahban et al. [2012]. Local restricted strong convexity condition were investigated in Loh and Wainwright [2015] and Mei et al. [2018] for some specific nonconvex loss functions. Those results, however, do not apply to our setting due to the estimated infinite-dimensional parameter.

Theorem 1 below presents non-asymptotic high-probability error bounds for any local sparse solution $\widehat{\boldsymbol{\beta}}$ that satisfies the penalized profiled estimation equation (9).

Theorem 1. *Assume conditions (A1)–(A5) in Section S2 of the online supplementary material are satisfied. Suppose* $\lambda = d_1 h^2$ *for some constant* $d_1 > 0$*, and* $d_0\left[\frac{s\log(p\vee n)}{n}\right]^{1/5} \leq h \leq d_0 n^{-1/6}$ *for some constant* $d_0 > 0$*. Then there exist universal positive constants* c_0 *and* c_1 *such that for any solution* $\widehat{\boldsymbol{\beta}}$ *in* $\mathbb{B}$*, we have*

$$||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_2 \leq \frac{6}{c_0}\lambda\sqrt{s}, \quad ||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_1 \leq \frac{24}{c_0}\lambda s,$$

with probability at least $1 - \exp(-c_1 \log p)$*, for all* n *sufficiently large.*

Remark 2. Theorem 1 shows that under some mild regularity conditions, local solutions of the profiled estimation equation (9) enjoy desirable estimation error rates, same as Lasso does for high-dimensional linear regression. For the purpose of inference, the initial estimator is not require to achieve perfect variable selection. The debiased estimator, however, can achieve the $n^{-1/2}$ rate for each individual coefficient, as we will show in Section 3.2. Carefully going through the proof of the theorem also reveals that the above error bounds hold uniformly for all $\boldsymbol{\beta}_0$ such that $||\boldsymbol{\beta}_0||_0 \le s$.

Remark 3. Based on Theorem 1, Lemmas A5–A6 in the online supplement establish the uniform convergence rates for the nonparametric estimator $\widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})$ and $\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})$, which are of independent interest. Under the assumptions of Theorem 1, we show that there exist universal positive constants c_0 and c_1 such that

$$P\Big(\max_{1\le i\le n}\sup_{\boldsymbol{\beta}\in\mathbb{B}}\big|\widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big| \ge c_0h^2\Big) \le \exp[-c_1\log(p\vee n)],$$
$$P\Big(\max_{1\le i\le n}\sup_{\boldsymbol{\beta}\in\mathbb{B}}\big|\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big| \ge c_0h\Big) \le \exp[-c_1\log(p\vee n)].$$

3.2 Theory for Inference

We first introduce some additional notation. Let $\widetilde{\boldsymbol{x}}_i = \boldsymbol{x}_i - \mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)$, and let $\widetilde{\boldsymbol{x}}_{i,-1}$ denote the $(p-1)$-subvector of $\widetilde{\boldsymbol{x}}_i$ that excludes its 1^{st} entry. Let $\boldsymbol{\Omega} = \mathrm{E}\big\{[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]^2\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T\big\}$. Assume the $(p-1)\times(p-1)$ matrix $\boldsymbol{\Omega}$ is positive definite and write its inverse $\boldsymbol{\Omega}^{-1} \triangleq \boldsymbol{\Theta} = (\boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_p)$. For $j = 2,\dots,p$, let $\boldsymbol{\Omega}_{-(j-1),-(j-1)} \in \mathbb{R}^{(p-2)\times(p-2)}$ be the submatrix of $\boldsymbol{\Omega}$ with its $(j-1)^{th}$ row and $(j-1)^{th}$ column removed; similarly $\boldsymbol{\Omega}_{-(j-1),(j-1)} \in \mathbb{R}^{p-1}$ denotes the $(j-1)^{th}$ column of $\boldsymbol{\Omega}$ with its $(j-1)^{th}$ entry removed. Note that $\boldsymbol{\Omega}_{-(j-1),-(j-1)}$ is positive definite. Define $\boldsymbol{d}_{0j} = (\boldsymbol{\Omega}_{-(j-1),-(j-1)})^{-1}\boldsymbol{\Omega}_{-(j-1),(j-1)}$, $s_j = ||\boldsymbol{d}_{0j}||_0$, $\widetilde{s} = \max_{2\le j\le p} s_j$ and $\tau_{0j}^2 = \boldsymbol{\Omega}_{(j-1),(j-1)} - \boldsymbol{d}_{0j}^T\boldsymbol{\Omega}_{-(j-1),(j-1)} = (\Theta_{(j-1),(j-1)})^{-1}$, $j = 2,\dots,p$.

Lemma 2 below establishes useful properties of the approximate inverse of $\nabla\boldsymbol{S}_n(\widehat{\boldsymbol{\beta}}, \widehat{G}, \widehat{\mathrm{E}})$, defined in Section 2.3.

Lemma 2. *Assume the conditions of Theorem 1 are satisfied. Let $\eta = d_2h$ for some positive constant $d_2 > 0$. If $\eta\widetilde{s} \le d_0$ and $d_0\big[\frac{s\log(p\vee n)}{n}\big]^{1/5} \le h \le d_0n^{-1/6}$ for some constant $d_0 > 0$,* *then there exist some universal positive constants d_2, c_0 and c_1 such that results (1)-(3) below hold uniformly in $j = 2, \ldots, p$, with probability at least $1 - \exp(-c_1 \log p)$ for all n sufficiently large:*

(1) $||\widehat{\boldsymbol{d}}_j - \boldsymbol{d}_{0j}||_2 \le \frac{8\sqrt{s_j}\eta}{\xi_2}$, *and* $||\widehat{\boldsymbol{d}}_j - \boldsymbol{d}_{0j}||_1 \le \frac{16 s_j \eta}{\xi_2}$;

(2) $|\tau_{0j}^2 - \widehat{\tau}_j^2| \le c_0\sqrt{s_j}\eta$, *and* $|\tau_{0j}^{-2} - \widehat{\tau}_j^{-2}| \le c_0\sqrt{s_j}\eta$;

(3) $||\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j||_2 \le c_0\sqrt{s_j}\eta$, *and* $||\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j||_1 \le c_0 s_j \eta$;

where $\xi_2 > 0$ is the smallest eigenvalue of $\boldsymbol{\Omega}$.

Lemma 2 requires $\tilde{s} = \max_{2 \le j \le p} s_j$ to be of order $O(h^{-1})$. For high-dimensional generalized linear models (Theorem 3.1, Van de Geer et al. [2014]), the corresponding sparsity constraint is $\tilde{s} = o(\sqrt{n/\log p})$. Our constrain is somewhat stricter due to the need to estimate the infinite-dimensional nuisance parameter. Building on Lemma 2, we prove the statistical property of the debiased estimator $\widetilde{\boldsymbol{\beta}}_{-1}$ defined in (10).

Theorem 2. *Assume the conditions of Lemma 2 are satisfied. Let $\Delta_{n,p} = sh^3\sqrt{n} + \widetilde{s}h\sqrt{\log p}$. Assume $\Delta_{n,p} = o(1)$ and $s\log(p \vee n) \le d_0 n h^5$ for some constant $d_0 > 0$. Then for all n sufficiently large,*

$$\sqrt{n}\big(\widetilde{\beta}_j - \beta_{0j}\big) = W_j + \Delta_j, \quad j = 2, \ldots, p,$$

with

$$W_j = n^{-1/2}\boldsymbol{e}_{j-1}^T \sum_{i=1}^n \widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\widetilde{\boldsymbol{x}}_{i,-1},$$
$$P\big(\max_{2\le j\le p} |\Delta_j| \ge c_0\Delta_{n,p}\big) \le \exp(-c_1 \log p),$$

where c_0, c_1 are universal positive constants, and $\boldsymbol{e}_{j-1}$ denotes the $(p-1)$-dimensional *vector with the $(j-1)^{th}$ entry being one and all the other entries equal to zero.*

Remark 4. Theorem 2 suggests that if we consider a lower-dimensional linear combination of coefficients $\boldsymbol{\alpha}^T\boldsymbol{\beta}_{0,-1}$, where $\boldsymbol{\alpha}$ is a $(p-1)$-dimensional nonzero vector of constants, then $\boldsymbol{\alpha}^T(\widetilde{\boldsymbol{\beta}}_{-1}-\boldsymbol{\beta}_{0,-1})$ has the asymptotic distribution $N(0, \boldsymbol{\alpha}^T\boldsymbol{\Theta}^T\boldsymbol{\Lambda}\boldsymbol{\Theta}\boldsymbol{\alpha})$ with $\boldsymbol{\Lambda} = \mathrm{E}\big\{[\widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]^2\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T\big\}$. The asymptotic covariance matrix resembles that in the literature on profiled estimation for index models in lower dimension, see Liang et al. [2010], Ma and He [2016], among other. The assumption $\Delta_{n,p}=o(1)$ is a sufficient condition for the remaining term of the linear approximation of $\sqrt{n}\big(\widetilde{\beta}_j-\beta_{0j}\big)$ to be uniformly negligible. It still allows p to grow at an exponential rate of n.

Remark 5. The proof of Theorem 2 is given in the online supplement. To build the theory, we show that

$$\begin{aligned}\sqrt{n}(\widetilde{\boldsymbol{\beta}}_{-1}-\boldsymbol{\beta}_{0,-1}) =& \sqrt{n}\widehat{\boldsymbol{\Theta}}^T\boldsymbol{S}_n(\boldsymbol{\beta}_0, G, \mathrm{E}) + \sqrt{n}(\boldsymbol{I}_{p-1}-\widehat{\boldsymbol{\Theta}}^T\boldsymbol{J}_1)(\widehat{\boldsymbol{\beta}}_{-1}-\boldsymbol{\beta}_{0,-1}) \\ & - \sqrt{n}\widehat{\boldsymbol{\Theta}}^T[\boldsymbol{S}_n(\widehat{\boldsymbol{\beta}}, \widehat{G}, \widehat{\mathrm{E}}) - \boldsymbol{S}_n(\boldsymbol{\beta}_0, G, \mathrm{E}) - \boldsymbol{J}_1(\widehat{\boldsymbol{\beta}}_{-1}-\boldsymbol{\beta}_{0,-1})] \\ \triangleq & \boldsymbol{A}_{n1}+\boldsymbol{A}_{n2}+\boldsymbol{A}_{n3},\end{aligned}$$

where $\boldsymbol{J}_1 = n^{-1}\sum_{i=1}^n[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2\widehat{\boldsymbol{x}}_{i,-1}\widehat{\boldsymbol{x}}_{i,-1}^T$ is the leading term in the approximation to $\nabla\boldsymbol{S}_n(\boldsymbol{\beta}_0, G, \mathrm{E})$. In the proof, we carefully justify that: (1) The $(j-1)^{th}$ component of $\boldsymbol{A}_{n1}$ can be approximated by W_j in the theorem, for $2\le j\le p$, (2) $P(||\boldsymbol{A}_{n2}||_\infty \ge c_0\Delta_{n,p}) \le \exp(-c_1\log p)$, and (3) $P(||\boldsymbol{A}_{n3}||_\infty \ge c_0\Delta_{n,p}) \le \exp(-c_1\log p)$, for some positive constants c_0 and c_1. Furthermore, to provide a deeper insight into the extension into the semiparametric setting, we consider the Gateaux functional derivative of the estimating function with respect to the infinite-dimensional nuisance parameters. Consider the functional $M(z;\boldsymbol{\beta},G,E) = [\widetilde{Y} - G(\boldsymbol{x}^T\boldsymbol{\beta}|\boldsymbol{\beta})]G^{(1)}(\boldsymbol{x}^T\boldsymbol{\beta}|\boldsymbol{\beta})[\boldsymbol{x}_{-1} - \mathrm{E}(\boldsymbol{x}_{-1}|\boldsymbol{x}^T\boldsymbol{\beta})]$, where $z = (A, X, \widetilde{Y})$ denotes a vector of random observations of the data. The Gateaux derivative of $M(z;\boldsymbol{\beta},G,E)$ at G in the direction $[\overline{G} - G]$ is defined as

$$\lim_{\tau\to 0} \frac{\mathbb{E}\big\{M(z;\boldsymbol{\beta},G+\tau(\overline{G}-G),E) - M(z;\boldsymbol{\beta},G,E)\big\}}{\tau}.$$

It is easy to see that this Gateaux derivative at G is zero when evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. Similarly, the Gateaux derivative with respect to E vanishes at the true value $\boldsymbol{\beta}_0$. This orthogonality behavior suggests the insensitivity of the estimating function to the infinite-dimensional nuisance parameters.

The following corollary establishes uniform validity of the marginal confidence intervals (15) introduced in Section 2.3.

Corollary 1. *Under the conditions of Theorem 2,*

$$\sup_{\boldsymbol{\beta}_0\in\mathbb{B}_0:||\boldsymbol{\beta}_0||_0\leq s}\ \max_{2\leq j\leq p}\ \sup_{\alpha\in(0,1)} \Big|P\Big(\big|\sqrt{n}(\widetilde{\beta}_j - \beta_{0j})\widehat{\Sigma}_{jj}^{-1/2}\big| \leq \Phi^{-1}(1-\alpha/2)\Big) - (1-\alpha)\Big| = o(1),$$

where $\widehat{\Sigma}_{jj}$ *denotes the* $(j-1)^{th}$ *diagonal entry of* $\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}})$ *defined in Section 2.3, and* $\Phi^{-1}(\cdot)$ *is the quantile function of* $N(0,1)$.

Finally, Theorem 3 below establishes the validity of the bootstrap procedure introduced in Section 2.3 for testing the group hypothesis (17). Given a group of variables $\mathcal{G} \subseteq \{2,\ldots,p\}$, the wild bootstrap test statistic is defined as $\sqrt{n}\max_{j\in\mathcal{G}}|\delta_j^*|$, where $\delta_j^* \triangleq n^{-1}\sum_{i=1}^n r_i\{\widetilde{Y}_i - \widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\}\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{x}}_{i,-1}^T\widehat{\boldsymbol{\theta}}_j$, and $r_1,\cdots,r_n$ are standard normal random variables that are independent of the data. Denote $\boldsymbol{r} = \{r_1,\cdots,r_n\}$, and let $\boldsymbol{w} = \{w_1,\cdots,w_n\}$ denote the random sample $w_i = (A_i,\boldsymbol{x}_i,\widetilde{Y}_i)$. Given $0<\alpha<1$, recall that the bootstrap critical value for a level-α test is defined as

$$c^*_{1-\alpha} = \inf\Big\{t \in \mathbb{R} : P\Big(\sqrt{n}\max_{j\in\mathcal{G}} |\delta^*_j| \le t \Big| \{w_i\}_{i=1}^n\Big) \ge 1-\alpha\Big\}. \tag{19}$$

Theorem 3. *Assume the conditions of Theorem 2 are satisfied. If* $\Delta_{n,p}\sqrt{\log p} = o(1)$, $h \ge d_0\big[\frac{s\log(p\vee n)}{n}\big]^{1/5}$ *for some constant* $d_0 > 0$, *and* $\sqrt{s}h\log^2 p = o(1)$, *then*

$$\sup_{\boldsymbol{\beta}_0\in\mathbb{B}_0:||\boldsymbol{\beta}_0||_0\le s}\ \sup_{\alpha\in(0,1)} \Big|P\Big(\sqrt{n}\max_{j\in\mathcal{G}}|\widetilde{\beta}_j - \beta_{0j}| \le c^*_{1-\alpha}(\mathcal{G})\Big) - (1-\alpha)\Big| = o(1).$$

Theorem 3 ensures that the multiplier bootstrap procedure is valid for the simultaneous testing problem (17). It is also *honest* in the sense of being valid uniformly over s-sparse models of the form (1). It does not require the nonzero components of $\boldsymbol{\beta}_0$ to be well-separated from zero. In particular, the multiple bootstrap procedure does not require the local solution of the profiled estimation to achieve perfect variable selection, which is usually unrealistic in practice.

4 Monte Carlo Studies

4.1 Algorithm for Estimation

To solve the penalized high-dimensional profiled estimating equation for the initial estimator $\widehat{\boldsymbol{\beta}}$, we extend the composite gradient algorithm (Nesterov [2007], Agarwal et al. [2012]) for high-dimensional M-estimator without nuisance parameters. A summary of the proposed algorithm is given in Algorithm 1 in Section S10.1 of the supplementary material.

Specifically, given a current estimator $\boldsymbol{\beta}^t = (1, (\boldsymbol{\beta}^t_{-1})^T)^T$ at step t, we update the estimate by

$$\boldsymbol{\beta}_{-1}^{t+1} = \mathop{\arg\min}_{\substack{\boldsymbol{\beta}_{-1}\in\mathbb{R}^{p-1} \\ ||\boldsymbol{\beta}_{-1}||_1\le\rho}} \Big\{ \frac{\gamma_u}{2}||\boldsymbol{\beta}_{-1} - \boldsymbol{\beta}_{-1}^t||_2^2 + [\boldsymbol{S}_n(\boldsymbol{\beta}^t, \widehat{G}, \widehat{\mathrm{E}})]^T(\boldsymbol{\beta}_{-1} - \boldsymbol{\beta}_{-1}^t) + \lambda||\boldsymbol{\beta}_{-1}||_1 \Big\}, \tag{20}$$

where γ_u is the step size, ρ is a positive constant such that $||\boldsymbol{\beta}_{0,-1}||_1 \le \rho$. An appealing practical property of the algorithm is that the update in step (20) can be done efficiently by the following formula:

$$\boldsymbol{\beta}_{-1}^{t+1} = T_s\Big(\boldsymbol{\beta}_{-1}^t - \frac{1}{\gamma_u}\boldsymbol{S}_n(\boldsymbol{\beta}^t, \widehat{G}, \widehat{\mathrm{E}}), \lambda\Big), \tag{21}$$

where the function $T_s(\boldsymbol{\beta}_{-1}, \lambda) = \Big(\mathrm{sgn}(\beta_j) * \max(|\beta_j| - \lambda, 0)\Big)_{j=2,\cdots,p}$ is the soft-threshold operator. Then to ensure the constraint $||\boldsymbol{\beta}_{-1}||_1 \le \rho$, we employ the projection method introduced in Duchi et al. [2008], which is described in Algorithm 2 in Section S10.1 of the online supplement.

In implementation, we choose the kernel function $K(\cdot)$ as the distribution function of the standard normal distribution. The bandwidth is set to be $h = 0.9n^{-1/6}\min\{\mathrm{std}(\boldsymbol{x}_i^T\boldsymbol{\beta}),$ $\mathrm{IQR}(\boldsymbol{x}_i^T\boldsymbol{\beta})/1.34\}$, as motivated by Silverman [1986], where "std" denotes the standard deviation, and "IQR" denotes the interquartile range. For the step-size parameter, inspired by Agarwal et al. [2012], we employ an expanding series for γ_u, which ensures that the stepsize diminishes during the update process. Given a set of candidate tuning parameters $\{\lambda_k\}$ and the corresponding estimators $\widehat{\boldsymbol{\beta}}_{\lambda_k}$, we employ 5-fold cross-validation to select the optimal tuning parameter λ by minimizing $\mathrm{MSE}(\lambda) = n^{-1}\sum_{i=1}^n\{\widetilde{Y}_i - \widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}_\lambda|\widehat{\boldsymbol{\beta}}_\lambda)\}^2$.

To obtain the debiased estimator $\widetilde{\boldsymbol{\beta}}$, the nodewise Dantzig estimator $\boldsymbol{d}_j(\widehat{\boldsymbol{\beta}}, \eta)$ in (11) is computed via linear programming, see details in Section S10.2 of the supplementary material.

4.2 Monte Carlo Results

We generate random data from the model $Y = (\boldsymbol{x}^T\boldsymbol{\eta})^2 + (A - \frac{1}{2})f_0(\boldsymbol{x}^T\boldsymbol{\beta}_0) + \epsilon$, where $\epsilon \sim N(0, 1)$, $A \sim$ Bernoulli(0.5), and $\boldsymbol{x}$ follows a p-dimensional multivariate normal distribution with mean zero and identity covariance matrix, $\boldsymbol{\eta} = (0.5, 0.5, -0.5, -0.5, 0, \cdots, 0)^T$, $\boldsymbol{\beta}_0 = (1, -1, -0.5, 0.4, -0.3, 0, \cdots, 0)^T$, and $f_0(u) = 20 * \{[1 + \exp(-u)]^{-1} - 0.5\}$. We consider $n = 300, 500$ and $p = 200, 800, 2000$ in the Monte Carlo experiment.

We first investigate the finite-sample performance of the penalized profiled semiparametric estimator in Section 2.2. Table 1 reports the average l_1- and l_2-estimation errors, the average number of false negatives (nonzero components incorrectly identified as zero) and false positives (zero components incorrectly identified as nonzero), with their standard errors in the parentheses, based on 500 simulation runs. Results in Table 1 demonstrate satisfactory performance of the profiled estimator for both the scenarios $p < n$ and $p > n$.

Table 1: Performance of the penalized profile least-squares estimator

n	p	l_1 error	l_2 error	False Negative	False Positive
	200	0.85 (0.02)	0.31 (0.00)	0.01 (0.01)	10.95 (0.32)
300	800	1.10 (0.03)	0.37 (0.00)	0.07 (0.01)	19.05 (1.13)
	2000	1.32 (0.03)	0.40 (0.00)	0.09 (0.01)	31.25 (1.57)
	200	0.58 (0.01)	0.22 (0.00)	0.00 (0.00)	9.30 (0.30)
500	800	0.79 (0.02)	0.27 (0.00)	0.00 (0.00)	17.39 (0.66)
	2000	0.94 (0.02)	0.31(0.00)	0.01 (0.00)	25.60 (1.18)

Next we investigate the wild bootstrap procedure introduced in Section 2.3 for testing the group hypothesis (17). We consider the following six different choices for the groups: $\mathcal{G}_1 = \{6, 7, 8, 9\}$, $\mathcal{G}_2 = \{5, 6, 7, 8, 9\}$, $\mathcal{G}_3 = \{4, 6, 7, 8, 9\}$, $\mathcal{G}_4 = \{4, 5, 6, 7, 8, 9\}$,

$\mathcal{G}_5 = \{3, 6, 7, 8, 9\}$ and $\mathcal{G}_6 = \{2, 6, 7, 8, 9\}$. Note that $\mathcal{G}_1$ consists of only zero entries in $\boldsymbol{\beta}_0$, while all the other groups include at least one non-zero elements. Table 2 summarizes the average Type I errors and powers for each scenario, based on 1000 Bootstrap samples and 500 simulation runs.

Table 2: Performance of the bootstrap procedure in Section 2.3 for simultaneous testing.

n	p	Type I error	Power				
		$\mathcal{G}_1$	$\mathcal{G}_2$	$\mathcal{G}_3$	$\mathcal{G}_4$	$\mathcal{G}_5$	$\mathcal{G}_6$
300	200	5.6%	96.4%	96.2%	97.8%	98.6%	100%
	800	5.4%	94.6%	97.6%	99.0%	99.6%	100%
	2000	3.2%	92.4%	96.8%	98.4%	99.0%	100%
500	200	4.4%	100%	100%	100%	100%	100%
	800	5.0%	99.6%	99.6%	100%	99.2%	100%
	2000	4.6%	98.8%	98.6%	99.0%	99.2%	100%

Table 2 indicates that type I errors are reasonable controlled for all scenarios. Power performance generally depends on the number and magnitudes of the nonzero components. The hypothesis corresponding to $\mathcal{G}_2$ represents a more challenging situation where the only non-zero element is -0.3, close to 0. The average powers for this case for different values of p are still over 90%.

Note that for inference, we need to estimate the approximate inverse of $\nabla S_n(\widehat{\boldsymbol{\beta}})$ which involves an additional tuning parameter η. We observe that the inference procedure is not overly sensitive to its choice and fix it at the value $\eta = 25h$ to save computational time. Alternatively, it can also be selected via cross-validation similarly as what has been done for λ selection. We provide additional simulation results in Section S10.3 of the online supplement, including investigation on the choice of η and comparing with alternative procedures for estimating the optimal value function.

5 A Real Data Example

We illustrate the application on a clinical data set introduced by Charbonnel et al. [2005]. This is a randomized, double-blind, parallel treatment arm, phase III clinical trial to compare the efficacy and safety of pioglitazone versus gliclazide on metabolic control in naive patients with Type 2 diabetes mellitus. This data set we consider contains information on clinical characteristics for 813 individuals with Type 2 diabetes. The patients were randomized into two treatment arms: pioglitazone (treatment 0) and gliclazide (treatment 1). Their glycosylated haemoglobin A_{1c} (HbA_{1c}) and fasting plasma glucose (FPG) levels were recorded every four weeks, up to week 52.

The primary efficacy endpoint is the change of HbA_{1c} from baseline to the last available post-treatment value. We consider the main effects of 22 baseline covariates and their two-way interactions in the model. The dimension of the model is over 250. In the analysis, we standardize the covariates to have mean zero and sample variance one.

We consider testing the significance of six different groups of variables. Table 3 summarizes these six different groups and their respective p-values, based on the bootstrap procedure in Section 2.3. The estimated coefficients are reported in Section S10.3 of the supplementary.

Table 3: Real data analysis: evaluation of the significance of different groups of variables

Group	Variables	$p-$value
1	HbA_{1c}, creatinine, BMI, waist circumference, HomaS	0.003
2	all variables in Group 1, all their two-way interactions, and their interactions with fasting insulin	0.011
3	HbA_{1c}, HomaS	< 0.001
4	BMI, creatinine, waist circumference,	0.242
5	LDL-C, total cholesterol, age, weight	0.494

Based on the scientific literature and suggestions from our clinical collaborators, fasting insulin is important for estimating the optimal treatment regime. We normalize its coefficient as 1 in our model. The first group includes the main effects of five characteristics, which are the baseline average levels for HbA_{1c}, creatinine, BMI, waist circumference and homeostatic model assessment insulin sensativity (HomaS). The variables in this group are those identified by diabetes experts to be potentially important for optimal treatment regime estimation. The bootstrap procedures suggests a significant p-value (0.003) for this group, which indicates that at one variable in this group is influential for making an optimal personalized decision in the choice of the two treatments. Group 2 augments Group 1 by including all the two-way interaction of these six characteristics (including fasting insulin), hence includes 20 variables in total. The estimated $p-$value is 0.011. Group 3 and Group 4 are subgroups of Group 1. The third group only includes two main effects: baseline HbA_{1c} and HomaS, while the fourth group includes the remaining three main effects. The estimated $p-$values suggest that the significant characteristics are among those in Group 3 rather than Group 4. Group 5 consists of four variables: the baseline average levels for the low-density lipoprotein cholesterol (LDL-C), total cholesterol, age and weight. This group of variables is of interest because Glucose and lipid metabolism are linked to each other in many ways [Parhofer, 2015]. Age and weight are also always taken into account for optimal treatment regime estimation. Our test suggests that Group 5 does not appear to be influential in optimal treatment recommendation.

6 Discussions

We propose a flexible semi-parametric approach for making honest simultaneous inference about the importance of a group of variables on optimal treatment regime estimation. We develop new statistical theory to overcome the challenges of nonconvexity, high dimensionality and infinite-dimensional nonparametric components.

In this paper, we focus on a randomized trial. For observation studies, let $\pi(\boldsymbol{x}) = P(A = 1|\boldsymbol{x})$ be the propensity score. Observing that $\mathrm{E}\{[A - \pi(\boldsymbol{x})]g(\boldsymbol{x})\} = 0$, we have

$$4[A_i - \pi(\boldsymbol{x}_i)]Y_i = 4[A_i - \pi(\boldsymbol{x}_i)]g(\boldsymbol{x}_i) + 4[A_i - \pi(\boldsymbol{x}_i)](A_i - 1/2)f_0(\boldsymbol{x}_i^T\boldsymbol{\beta}_0) + 4[A_i - \pi(\boldsymbol{x}_i)]\epsilon_i.$$

Let $\widetilde{Y}_i = 4[A_i - \pi(\boldsymbol{x}_i)]Y_i$, $\widetilde{\epsilon}_i = 4[A_i - \pi(\boldsymbol{x}_i)][\epsilon_i + g(\boldsymbol{x}_i)]$, then we have

$$\mathrm{E}\widetilde{Y}_i = 4[A_i - \pi(\boldsymbol{x}_i)](A_i - 1/2)f_0(\boldsymbol{x}_i^T\boldsymbol{\beta}_0).$$

Denote $G(t|\boldsymbol{\beta}) = \mathrm{E}(\widetilde{Y}|\boldsymbol{x}^T\boldsymbol{\beta} = t) = 2\mathrm{E}\{[A - \pi(\boldsymbol{x})](2A - 1)f_0(\boldsymbol{x}^T\boldsymbol{\beta}_0)|\boldsymbol{x}^T\boldsymbol{\beta} = t\}$, $G^{(1)}(t|\boldsymbol{\beta}) = \frac{d}{dt}G(t|\boldsymbol{\beta})$, and define $\widehat{G}(t|\boldsymbol{\beta})$, $\widehat{G}^{(1)}(t|\boldsymbol{\beta})$ similarly as in Section 2.2. Assume $\pi(\boldsymbol{x}) = P(A = 1|\boldsymbol{x})$ can be modeled as $\pi(\boldsymbol{x}, \boldsymbol{\xi})$, where $\boldsymbol{\xi}$ is a finite-dimensional parameter. Let $\widehat{\boldsymbol{\xi}}$ be an estimate of $\boldsymbol{\xi}$, such as the one based on the regularized logistic regression. Define the profiled semiparametric estimating function $\boldsymbol{S}_n(\boldsymbol{\beta}, \widehat{G}, \widehat{\mathrm{E}}, \widehat{\boldsymbol{\xi}}) = -n^{-1}\sum_{i=1}^n\{4[A_i - \pi(\boldsymbol{x}_i, \widehat{\boldsymbol{\xi}})]Y_i - \widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\}\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})[\boldsymbol{x}_{i,-1} - \widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})]$. We then estimate $\boldsymbol{\beta}_0$ through the following penalized semiparametric profiled estimating equation $\boldsymbol{S}_n(\boldsymbol{\beta}, \widehat{G}, \widehat{\mathrm{E}}, \widehat{\boldsymbol{\xi}}) + \lambda\widehat{\boldsymbol{\kappa}} = \mathbf{0}$. Promising numerical performance of this estimator is reported in Section S10.3 of the supplementary. Our approach can still be applied to investigate the theory but is it more complex due to the additional nuisance parameter. We will explore the complete theory for the above estimator in the future work. Alternative approaches that can potentially be extended to our setting include Nie and Wager [2020], Künzel et al. [2018], among ohers.

Our approach for high-dimensional inference generalizes the "inverting KKT condition" technique in Van de Geer et al. [2014]. An alternative approach, which is more suitable if one is interested in some targeted lower-dimensional parameter is based on the idea of orthogonalization, see for example Belloni et al. [2015], Ning et al. [2017], Chernozhukov et al. [2018]. In contrast, our approach is able to achieve debiasing for the p-dimensional coefficient vector simultaneously. The main idea of the orthogonalization approach is to construct a lower-dimensional estimating equation which is locally insensitive to the nuisance parameters. The construction of such a lower-dimensional moment condition is non-trivial for high-dimensional semiparametric setting, particularly for index model, where the challenge of bundled parameter arises.

References

A. Agarwal, S. Negahban, and M. J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics*, 40(5):2452–2482, 2012.

A. Belloni, V. Chernozhukov, and K. Kato. Uniform post-selection inference for least absolute deviation regression and other z-estimation problems. *Biometrika*, 102(1):77–94, 2015.

T. T. Cai, Z. Guo, et al. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics*, 45(2):615–646, 2017.

R. J. Carroll, J. Fan, I. Gijbels, and M. P. Wand. Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92(438):477–489, 1997.

B. Chakraborty and E. E. Moodie. *Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine*. Springer Science & Business Media, 2013.

B. Chakraborty, E. B. Laber, and Y. Zhao. Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. *Biometrics*, 69(3):714–723, 2013.

B. Chakraborty, E. B. Laber, and Y.-Q. Zhao. Inference about the expected performance of a data-driven dynamic treatment regime. *Clinical Trials*, 11(4):408–417, 2014.

B. H. Charbonnel, D. R. Matthews, G. Schernthaner, M. Hanefeld, P. Brunetti, and on behalf of the QUARTET Study Group. A long-term comparison of pioglitazone and gliclazide in patients with type 2 diabetes mellitus: a randomized, double-blind, parallel-group comparison trial. *Diabetic Medicine*, 22(4):399–405, 2005.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Y. Ding and B. Nan. A sieve M-theorem for bundled parameters in semiparametric models, with application to the efficient estimation in a linear model for censored data. *The Annals of Statistics*, 39(6):3032–3061, 2011.

N. Duan and K.-C. Li. Slicing regression: A link-free regression method. *The Annals of Statistics*, 19(2):505–530, 1991.

J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine Learning*, pages 272–279, 2008.

T. Gueuning and G. Claeskens. Confidence intervals for high-dimensional partially linear single-index models. *Journal of Multivariate Analysis*, 149:13–29, 2016.

H. Ichimura. Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58(1-2):71–120, 1993.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

X. J. Jeng, W. Lu, H. Peng, et al. High-dimensional inference for personalized treatment decision. *Electronic Journal of Statistics*, 12(1):2074–2089, 2018.

M. R. Kosorok and E. E. Moodie. *Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine.* ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA., 2016.

S. R. Künzel, S. J. S. Walter, and J. S. Sekhon. Causaltoolbox—estimator stability for heterogeneous treatment effects. *arXiv: Methodology*, 2018.

E. B. Laber, D. J. Lizotte, M. Qian, W. E. Pelham, and S. A. Murphy. Dynamic treatment regimes: Technical challenges and applications. *Electronic Journal of Statistics*, 8(1): 1225 – 1272, 2014.

K.-C. Li. Honest Confidence Regions for Nonparametric Regression. *The Annals of Statistics*, 17(3):1001–1008, 1989.

H. Liang, X. Liu, R. Li, and C.-L. Tsai. Estimation and testing for partially linear single-index models. *The Annals of Statistics*, 38(6):3811–3836, 2010.

Q. Lin, Z. Zhao, and J. S. Liu. Sparse sliced inverse regression via lasso. *Journal of the American Statistical Association*, 114(528):1726–1739, 2019.

P.-L. Loh and M. J. Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16: 559–616, 2015.

A. R. Luedtke and M. J. van der Laan. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *The Annals of Statistics*, 44(2):713–742, 2016.

S. Ma and X. He. Inference for single-index quantile regression models with profile optimization. *The Annals of Statistics*, 44(3):1234–1268, 2016.

Y. Ma and L. Zhu. Doubly robust and efficient estimators for heteroscedastic partially linear single-index models allowing high dimensional covariates. *Journal of the Royal Statistical Society: Series B*, 75(2):305–322, 2013.

I. W. McKeague and M. Qian. An adaptive resampling test for detecting the presence of significant predictors. *Journal of the American Statistical Association*, 110(512):1422–1433, 2015.

S. Mei, Y. Bai, and A. Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

E. E. Moodie and T. S. Richardson. Estimating optimal dynamic regimes: Correcting bias under the null. *Scandinavian Journal of Statistics*, 37(1):126–146, 2010.

S. A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B*, 65(2):331–366, 2003.

S. N. Negahban, P. Ravikumar, M. J. Wainwright, B. Yu, et al. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Y. Nesterov. Gradient methods for minimizing composite objective function. Core discussion papers, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2007.

M. Neykov, J. S. Liu, and T. Cai. L1-regularized least squares for support recovery of high dimensional single index models with gaussian designs. *The Journal of Machine Learning Research*, 17(1):2976–3012, 2016.

J. Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4):465–472, 1990.

X. Nie and S. Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 09 2020.

Y. Ning, H. Liu, et al. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45(1):158–195, 2017.

K. G. Parhofer. Interaction between glucose and lipid metabolism: more than diabetic dyslipidemia. *Diabetes & Metabolism Journal*, 39(5):353–362, 2015.

B. M. Pötscher. Confidence sets based on sparse estimators are necessarily large. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 1–18, 2009.

J. L. Powell, J. H. Stock, and T. M. Stoker. Semiparametric estimation of index coefficients. *Econometrica*, pages 1403–1430, 1989.

Z. Qi, Y. Liu, et al. D-learning to estimate optimal individual treatment rules. *Electronic Journal of Statistics*, 12(2):3601–3638, 2018.

M. Qian and S. A. Murphy. Performance guarantees for individualized treatment rules. *The Annals of Statistics*, 39(2):1180–1210, 2011.

P. Radchenko. High dimensional single index models. *Journal of Multivariate Analysis*, 139:266–282, 2015.

J. Robins, M. Hernan, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11:550–560, 2000.

D. B. Rubin. Bayesian inference for causal effects: the role of randomization. *The Annals of Statistics*, 6:34–58, 1978.

C. Shi, R. Song, W. Lu, and R. Li. Statistical inference for high-dimensional models via recursive online-score estimation. *Journal of the American Statistical Association*, 2020.

B. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1986.

R. Song, W. Wang, D. Zeng, and M. Kosorok. Penalized q-learning for dynamic treatment regimens. *Statistica Sinica*, 25:901–920, 2015.

L. Tian, A. A. Alizadeh, A. J. Gentles, and R. Tibshirani. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532, 2014.

S. Van de Geer, P. Bühlmann, Y. Ritov, R. Dezeure, et al. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42 (3):1166–1202, 2014.

J.-L. Wang, L. Xue, L. Zhu, Y. S. Chong, et al. Estimation for a partial-linear single-index model. *The Annals of Statistics*, 38(1):246–274, 2010.

L. Wang, Y. Zhou, R. Song, and B. Sherwood. Quantile-optimal treatment regimes. *Journal of the American Statistical Association*, 113(523):1243–1254, 2018.

C. J. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8(3-4):279–292, 1992.

Y. Xia, H. Tong, and W. K. Li. On extended partially linear single-index models. *Biometrika*, 86(4):831–842, 1999.

Z. Yang, K. Balasubramanian, and H. Liu. High-dimensional non-gaussian single index models via thresholded score function estimation. In *International Conference on Machine Learning*, pages 3851–3860, 2017.

Y. Yu and D. Ruppert. Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, 97(460):1042–1054, 2002.

B. Zhang, A. A. Tsiatis, E. B. Laber, and M. Davidian. A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018, 2012.

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B*, 76(1): 217–242, 2014.

X. Zhang and G. Cheng. Simultaneous inference for high-dimensional linear models. *Journal of the American Statistical Association*, 112(518):757–768, 2017.

Y. Zhao, D. Zeng, A. J. Rush, and M. R. Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.

Y. Zhao, D. Zeng, E. B. Laber, and M. R. Kosorok. New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*, 110:583–598, 2015.

L. Zhu and L. Xue. Empirical likelihood confidence regions in a partially linear single-index model. *Journal of the Royal Statistical Society: Series B*, 68(3):549–570, 2006.

W. Zhu, D. Zeng, and R. Song. Proper inference for value function in high-dimensional q-learning for dynamic treatment regimes. *Journal of the American Statistical Association*, 114(527):1404–1417, 2019.

Y. Zhu and J. Bradic. Linear hypothesis testing in dense high-dimensional linear models. *Journal of the American Statistical Association*, 113(524):1583–1600, 2018.

Supplementary Material for *Model-Assisted Uniformly Honest Inference for Optimal Treatment Regimes in High Dimension*

Yunan Wu, Lan Wang and Haoda Fu

University of Texas at Dallas, University of Miami and Eli Lilly and Company

Section S1 of the supplementary material summarizes all the commonly used notation in the proof. Section S2 summaries the regularity conditions and presents some discussions on these conditions. Section S3 presents the technical lemmas used in the proof. Section S4 and Section S5 of the supplementary material provide the proofs of the theoretical results in Section 3.1 and 3.2 of the main paper, respectively. Section S6 presents the proofs of the technical lemmas in Section S3. Section S7 provides additional technical details. Section S8 introduces the identification conditions for the single index models mentioned in Assumption (A1)-(c). Section S9 provides examples for verifying the regularity conditions. Section S10 presents the algorithms introduced in Section 4.1 of the main paper and some additional numerical results.

[1]Yunan Wu is Assistant Professor, Department of Mathematical Sciences, University of Texas at Dallas. Emails: yunan.wu@utdallas.edu. Lan Wang is Professor, Department of Management Science, University of Miami. Emails: lanwang@mbs.miami.edu. Wang and Wu's research was partly supported by NSF DMS-1952373 and NSF OAC-1940160. Dr. Haoda Fu is Research Fellow, Enterprise Lead for Machine Learning and AI, Eli Lilly and Company. Email: fu_haoda@lilly.com. Wang and Wu's research was supported by NSF DMS-1712706. The authors are grateful to the referees, the associate editor and the Co-editor for their valuable comments, which have significantly improved the paper.

S1 Review of some useful notation

We collect below some notation introduced in the main paper for easy reference in the proof. First, recall model (1) in the main paper is

$$Y_i = g(\boldsymbol{x}_i) + (A_i - 1/2)f_0(\boldsymbol{x}_i^T\boldsymbol{\beta}_0) + \epsilon_i, \quad i = 1, \ldots, n.$$

Recall $G(t|\boldsymbol{\beta}) = \mathrm{E}\{f_0(\boldsymbol{x}^T\boldsymbol{\beta}_0)|\boldsymbol{x}^T\boldsymbol{\beta} = t\}$ and $G^{(1)}(t|\boldsymbol{\beta}) = \frac{d}{dt}G(t|\boldsymbol{\beta})$. Note that $G(t|\boldsymbol{\beta}_0) = f_0(t)$ and $G^{(1)}(t|\boldsymbol{\beta}_0) = f_0'(t)$. We assume that $\boldsymbol{\beta}_0 \in \mathbb{B}_0 = \{\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T : \beta_1 = 1\}$.

Denote $\boldsymbol{x}_i = (x_{i,1}, \boldsymbol{x}_{i,-1}^T)^T$ and $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_{-1}^T)^T$. The estimated profiled score function is

$$\boldsymbol{S}_n(\boldsymbol{\beta}, \widehat{G}, \widehat{\mathrm{E}}) = -\, n^{-1}\sum_{i=1}^{n}\big[\widetilde{Y}_i - \widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big[\boldsymbol{x}_{i,-1} - \widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})\big],$$

where $\widetilde{Y}_i = 2(2A_i - 1)Y_i$, $\widetilde{\epsilon}_i = 2(2A_i - 1)[\epsilon_i + g(\boldsymbol{x}_i)]$, and

$$\widehat{G}(t|\boldsymbol{\beta}) = \sum_{i=1}^{n} W_{ni}(t, \boldsymbol{\beta})\widetilde{Y}_i, \quad \widehat{G}^{(1)}(t|\boldsymbol{\beta}) = \sum_{i=1}^{n} W_{ni}^{(1)}(t, \boldsymbol{\beta})\widetilde{Y}_i, \quad \widehat{\mathrm{E}}(\boldsymbol{x}_{-1}|\boldsymbol{x}^T\boldsymbol{\beta} = t) = \sum_{i=1}^{n} W_{ni}(t, \boldsymbol{\beta})\boldsymbol{x}_{i,-1},$$

with $K_h(z) = h^{-1}K(z/h)$, $W_{ni}(t, \boldsymbol{\beta}) = \frac{K_h(t - \boldsymbol{x}_i^T\boldsymbol{\beta})}{\sum_{j=1}^{n} K_h(t - \boldsymbol{x}_j^T\boldsymbol{\beta})}$, and $W_{ni}^{(1)}(t, \boldsymbol{\beta}) = \frac{d}{dt}W_{ni}(t, \boldsymbol{\beta})$. Note that to estimate $\widehat{G}(\boldsymbol{x}_j^T\boldsymbol{\beta}|\boldsymbol{\beta})$, $\widehat{G}^{(1)}(\boldsymbol{x}_j^T\boldsymbol{\beta}|\boldsymbol{\beta})$ and $\widehat{\mathrm{E}}(\boldsymbol{x}_{j,-1}|\boldsymbol{x}_j^T\boldsymbol{\beta})$, we employ the leave-one-out estimators.

The debiased estimator is defined as

$$\widetilde{\boldsymbol{\beta}}_{-1} = \widehat{\boldsymbol{\beta}}_{-1} - \widehat{\boldsymbol{\Theta}}^T\boldsymbol{S}_n(\widehat{\boldsymbol{\beta}}, \widehat{G}, \widehat{\mathrm{E}}),$$

where the $(p-1) \times (p-1)$ matrix $\widehat{\boldsymbol{\Theta}} = (\widehat{\boldsymbol{\theta}}_2, ..., \widehat{\boldsymbol{\theta}}_p)$ is an approximation to the inverse of $\nabla \boldsymbol{S}_n(\widehat{\boldsymbol{\beta}}, \widehat{G}, \widehat{\mathrm{E}})$, where ∇ denotes the gradient with respect to the components of $\boldsymbol{\beta}_{-1}$. Given an initial estimator $\widehat{\boldsymbol{\beta}}$, to compute $\widehat{\boldsymbol{\theta}}_j$, let

$$\widehat{\boldsymbol{d}}_j \triangleq \boldsymbol{d}_j(\widehat{\boldsymbol{\beta}}, \eta) = \underset{\boldsymbol{v} \in \mathbb{R}^{p-2}}{\arg\min} ||\boldsymbol{v}||_1 \text{ subject to } \Big|\Big|n^{-1} \sum_{i=1}^{n} \big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\big]^2 (\widehat{x}_{i,j} - \widehat{\boldsymbol{x}}_{i,-j*}^T \boldsymbol{v}) \widehat{\boldsymbol{x}}_{i,-j*}\Big|\Big|_\infty \leqslant \eta,$$

for some $\eta > 0$, $j = 2, \cdots, p$, where $|| \cdot ||_\infty$ denotes the infinity norm of a vector, $\widehat{\boldsymbol{x}}_i = \boldsymbol{x}_i - \widehat{\mathrm{E}}(\boldsymbol{x}_i | \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}})$, $\widehat{x}_{i,j}$ denotes the j^{th} entry of the vector $\widehat{\boldsymbol{x}}_i$, $\widehat{\boldsymbol{x}}_{i,-1}$ denotes the $(p-1)$-subvector of $\widehat{\boldsymbol{x}}_i$ that excludes its 1^{st} entry, and $\widehat{\boldsymbol{x}}_{i,-j*}$ denotes the $(p-2)$-subvector of $\widehat{\boldsymbol{x}}_i$ that excludes both its 1^{st} and j^{th} entries. Furthermore, for $j = 2, \cdots, p$, let

$$\begin{aligned}
\boldsymbol{\phi}_j(\widehat{\boldsymbol{\beta}}, \eta) &= \Big(-\big(\boldsymbol{d}_j(\widehat{\boldsymbol{\beta}}, \eta)\big)^T_{1:(j-2)}, 1, -\big(\boldsymbol{d}_j(\widehat{\boldsymbol{\beta}}, \eta)\big)^T_{(j-1):(p-2)} \Big)^T, \\
\widehat{\tau}_j^2 \triangleq \tau_j^2(\widehat{\boldsymbol{\beta}}, \eta) &= n^{-1} \sum_{i=1}^{n} \big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\big]^2 \widehat{x}_{i,j} \widehat{\boldsymbol{x}}_{i,-1}^T \boldsymbol{\phi}_j(\widehat{\boldsymbol{\beta}}, \eta), \\
\widehat{\boldsymbol{\theta}}_j \triangleq \boldsymbol{\theta}_j(\widehat{\boldsymbol{\beta}}, \eta) &= \tau_j^{-2}(\widehat{\boldsymbol{\beta}}, \eta) \boldsymbol{\phi}_j(\widehat{\boldsymbol{\beta}}, \eta).
\end{aligned}$$

Then we can define the matrix $\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}})$ as below:

$$\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}) \triangleq \widehat{\boldsymbol{\Theta}}^T \Big\{ \frac{1}{n} \sum_{i=1}^{n} \big[\widetilde{Y}_i - \widehat{G}(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\big]^2 [\widehat{G}^{(1)}(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2 \widehat{\boldsymbol{x}}_{i,-1} \widehat{\boldsymbol{x}}_{i,-1}^T \Big\} \widehat{\boldsymbol{\Theta}}.$$

Define the matrix $\boldsymbol{\Omega} = \mathrm{E}\big\{[G^{(1)}(\boldsymbol{x}_i^T \boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]^2 \widetilde{\boldsymbol{x}}_{i,-1} \widetilde{\boldsymbol{x}}_{i,-1}^T\big\}$, with $\widetilde{\boldsymbol{x}}_{i,-1} = \boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1} | \boldsymbol{x}_i^T \boldsymbol{\beta}_0)$, and its inverse $\boldsymbol{\Omega}^{-1} \triangleq \boldsymbol{\Theta} = (\boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_p)$. For $j = 2, \ldots, p$, let $\boldsymbol{\Omega}_{-(j-1),-(j-1)} \in \mathbb{R}^{(p-2)\times(p-2)}$ be the submatrix of $\boldsymbol{\Omega}$ with its $(j-1)^{th}$ row and $(j-1)^{th}$ column removed; similarly $\boldsymbol{\Omega}_{-(j-1),(j-1)} \in \mathbb{R}^{p-2}$ denotes the $(j-1)^{th}$ column of $\boldsymbol{\Omega}$ with its $(j-1)^{th}$ entry removed. Define $\boldsymbol{d}_{0j} = (\boldsymbol{\Omega}_{-(j-1),-(j-1)})^{-1} \boldsymbol{\Omega}_{-(j-1),(j-1)}$, $s_j = ||\boldsymbol{d}_{0j}||_0$, $\widetilde{s} = \max_{2 \leqslant j \leqslant p} s_j$ and $\tau_{0j}^2 = \Omega_{(j-1),(j-1)} - \boldsymbol{d}_{0j}^T \boldsymbol{\Omega}_{-(j-1),(j-1)} = \big[\Theta_{(j-1),(j-1)}\big]^{-1}$, $\boldsymbol{\phi}_{0j} = \tau_{0j}^2 \boldsymbol{\theta}_j$. Denote $\mathbb{K}(p, s_0) = \{\boldsymbol{v} \in \mathbb{R}^p : ||\boldsymbol{v}||_2 \leqslant 1, ||\boldsymbol{v}||_0 \leqslant s_0\}$ for any integers p and s_0. Finally, recall $s = ||\boldsymbol{\beta}_0||_0$, $\mathbb{B}_1 = \big\{\boldsymbol{\beta} \in \mathbb{B} : ||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_2 \leqslant c_0 \sqrt{s} h^2, ||\boldsymbol{\beta}||_0 \leqslant ks\big\}$, where $\mathbb{B} = \big\{\boldsymbol{\beta} \in \mathbb{B}_0 : ||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_2 \leqslant r, ||\boldsymbol{\beta}||_0 \leqslant ks\big\}$.

For any $p-$dimensional vector $\boldsymbol{v} = (v_1, \cdots, v_p)$, we denote $||\boldsymbol{v}||_1 = \sum_{j=1}^p |v_j|$, $||\boldsymbol{v}||_2 = \sqrt{\sum_{j=1}^p |v_j|^2}$, and $||\boldsymbol{v}||_\infty = \max_{1 \leqslant j \leqslant p} |v_j|$. For any matrix $\boldsymbol{A} = (a_{ij}) \in \mathbb{R}^{p_1 \times p_2}$, where p_1, p_2

are two arbitrary integers, we denote $||\boldsymbol{A}||_\infty = \max_{1\leqslant i\leqslant p_1, 1\leqslant j\leqslant p_2} |a_{ij}|$.

S2 Regularity Conditions

We define some notation first. Given any square matrix A, $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote the largest and the smallest eigenvalues of A, respectively. Let $\mathcal{V}_1 = \{\boldsymbol{v} \in \mathbb{R}^{p-1} : ||\boldsymbol{v}||_2 = 1, ||\boldsymbol{v}||_0 \leqslant 2ks\}$, where $k >$ is a positive integer. Let $\mathrm{E}^{(j)}(\boldsymbol{x}_{-1}|\boldsymbol{x}^T\boldsymbol{\beta} = t)$ denote the derivatives of $\mathrm{E}(\boldsymbol{x}_{-1}|\boldsymbol{x}^T\boldsymbol{\beta} = t)$ with respect to t, $j = 1, 2$. Let $\mathrm{E}^{(1)}(\boldsymbol{x}_{-1}\boldsymbol{x}_{-1}^T|\boldsymbol{x}^T\boldsymbol{\beta} = t)$ denote the derivative of $\mathrm{E}(\boldsymbol{x}_{-1}\boldsymbol{x}_{-1}^T|\boldsymbol{x}^T\boldsymbol{\beta} = t)$ with respect to t.

We state below a set of regularity conditions, followed by Remarks (a)–(c) to discuss these conditions.

(A1) (a) The distributions of $\boldsymbol{x} \in \mathbb{R}^p$ and ϵ are sub-Gaussian with variance proxy σ_x^2 and σ_ϵ^2, respectively, where $p \geqslant 2$.

(b) The function $f_0(\cdot)$ satisfies $\mathrm{E}\{[f_0'(\boldsymbol{x}^T\boldsymbol{\beta}_0)]^2\} = a^2 > 0$ and $\max_{1\leqslant i\leqslant n} |f_0'(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)| \leqslant b$ for some positive constants a and b, where f_0' denotes its first derivative, and $||\boldsymbol{\beta}_0||_2$ is bounded. Its second derivative $f_0''(z)$ and third derivative $f_0'''(z)$ are bounded for $z \in \mathbb{R}$.

(c) The lower-dimensional true model $f_0(\boldsymbol{x}^T\boldsymbol{\beta}_0)$ satisfies the identifiability conditions for the classical single index models (e.g., Ichimura [1993], Horowitz [2012], see Section S8 of the online supplement for details).

(d) The main effect $g(\cdot)$ satisfies $P\big(\max_{1\leqslant i\leqslant n} |g(\boldsymbol{x}_i)| \leqslant M\big) = 1$ for some positive constant M.

(A2) (a) There exist some positive constants M, ξ_0, ξ_1, ξ_2, ξ_3 and ξ_4 such that we have $\inf_{\boldsymbol{v}\in\mathcal{V}_1} \boldsymbol{v}^T\mathrm{E}\big[\mathrm{Cov}(\boldsymbol{x}_{-1}|\boldsymbol{x}^T\boldsymbol{\beta}_0)\big]\boldsymbol{v} \geqslant \xi_0$, $\lambda_{\max}(\mathrm{E}\big[\mathrm{Cov}(\boldsymbol{x}_{-1}|\boldsymbol{x}^T\boldsymbol{\beta}_0)\big]) \leqslant \xi_1$, $\lambda_{\min}(\boldsymbol{\Omega}) \geqslant \xi_2$, and $\lambda_{\max}(\mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T)) \leqslant \xi_3$. Also, $\sup_{\boldsymbol{\beta}\in\mathbb{B}} n^{-1}\sum_{i=1}^n \big[\lambda_{\max}\big(\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i^T|\boldsymbol{x}_i^T\boldsymbol{\beta})\big)\big]^2 \leqslant \xi_4$

and $\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\lambda_{\max}\big(\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i^T|\boldsymbol{x}_i^T\boldsymbol{\beta})\big)\leqslant M\log(p\vee n)$, for all n sufficiently large.

(b) $\mathrm{E}(\boldsymbol{x}_{-1}|\boldsymbol{x}^T\boldsymbol{\beta}=t)$ is twice-differentiable with respect to t, and $\mathrm{E}(\boldsymbol{x}_{-1}\boldsymbol{x}_{-1}^T|\boldsymbol{x}^T\boldsymbol{\beta}=t)$ is differentiable with respect to t. There exists some positive constant M such that for any $\boldsymbol{\eta}\in\mathbb{R}^{p-1}$, $\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}}|\mathrm{E}^{(1)}(\boldsymbol{x}_{i,-1}^T\boldsymbol{\eta}|\boldsymbol{x}_i^T\boldsymbol{\beta})|\leqslant M||\boldsymbol{\eta}||_2$, and $\sup_{|t|\leqslant 2||\boldsymbol{\beta}_0||_2\sigma_x\sqrt{\log(p\vee n)}}\sup_{\boldsymbol{\beta}\in\mathbb{B}}|\mathrm{E}^{(2)}(\boldsymbol{x}_{-1}^T\boldsymbol{\eta}|\boldsymbol{x}^T\boldsymbol{\beta}=t)|\leqslant M||\boldsymbol{\eta}||_2$.
Furthermore, $\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\{\big|\mathrm{E}^{(1)}[(\boldsymbol{x}_{i,-1}^T\boldsymbol{\eta})^2]|\boldsymbol{x}_i^T\boldsymbol{\beta}\big|\leqslant M||\boldsymbol{\eta}||_2^2\sqrt{\log(p\vee n)}$, for all n sufficiently large.

(c) For some positive constant C,

$$\begin{aligned}&\sup_{\boldsymbol{v}\in\mathbb{K}(p,2ks+\widetilde{s})}\big|[\mathrm{E}(\boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\beta}_1)-\mathrm{E}(\boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\beta}_2)]^T\boldsymbol{v}\big|\\ \leqslant & C\Big(|\boldsymbol{x}^T\boldsymbol{\beta}_1-\boldsymbol{x}^T\boldsymbol{\beta}_2|+\max(|\boldsymbol{x}^T\boldsymbol{\beta}_1|,|\boldsymbol{x}^T\boldsymbol{\beta}_2|)||\boldsymbol{\beta}_1-\boldsymbol{\beta}_2||_2\Big),\end{aligned}$$

for any $\boldsymbol{\beta}_1,\boldsymbol{\beta}_2\in\mathbb{B}$, $\boldsymbol{x}\in\mathbb{R}^p$, where $\mathbb{K}(p,2ks+\widetilde{s})=\{\boldsymbol{v}\in\mathbb{R}^p:||\boldsymbol{v}||_2\leqslant 1,||\boldsymbol{v}||_0\leqslant 2ks+\widetilde{s}\}$, and $\widetilde{s}=\max_{1\leqslant j\leqslant p}||\boldsymbol{d}_{0j}||_0$.

(A3) The kernel function $K(\cdot)$ is nonnegative, symmetric about 0, and twice differentiable and bounded on the real line. The function $K(\cdot)$ and its derivatives $K'(\cdot)$, $K''(\cdot)$ are all Lipschitz on the real line. Furthermore, $\lim_{|\nu|\to\infty}K(\nu)=0$, $\int_{-\infty}^{\infty}K(\nu)d\nu=1$, $\int_{-\infty}^{\infty}\nu K'(\nu)d\nu=-1$, and $\int_{-\infty}^{\infty}\nu^2K''(\nu)d\nu=2$. For any integer $0\leqslant i\leqslant 4$ $\int|\nu^iK(\nu)|d\nu<\infty$; for integer $0\leqslant i\leqslant 3$, $\int|\nu^iK'(\nu)|d\nu<\infty$; for integer $0\leqslant i\leqslant 2$, $\int|\nu^iK''(\nu)|d\nu<\infty$.

(A4) Let $f_{\boldsymbol{\beta}}(\cdot)$ denote the density function of $\boldsymbol{x}^T\boldsymbol{\beta}$. Suppose that $f_{\boldsymbol{\beta}}(\cdot)$ is twice differentiable, and $f_{\boldsymbol{\beta}}(\cdot)$, $f'_{\boldsymbol{\beta}}(\cdot)$, $f''_{\boldsymbol{\beta}}(\cdot)$ are all bounded on the real line. Furthermore, for some positive constant M, $P(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}}f_{\boldsymbol{\beta}}^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta})\leqslant M)=1$. .

(A5) (a) For any $\boldsymbol{\beta}\in\mathbb{B}$ and $t\in\mathbb{R}$, $G(t|\boldsymbol{\beta})=\mathrm{E}\{f_0(\boldsymbol{x}^T\boldsymbol{\beta}_0)|\boldsymbol{x}^T\boldsymbol{\beta}=t\}$ is twice differentiable with respect to t. Its first derivative satisfies $P(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}}|G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})|\leqslant$

$b) = 1$, for some positive constant b. Its second derivative $G^{(2)}(t|\boldsymbol{\beta})$ is bounded.

(b) $G^{(1)}(t|\boldsymbol{\beta})$ satisfies

$$n^{-1}\sum_{i=1}^{n}\left[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right]^2 \leqslant c_1||\boldsymbol{\beta}-\boldsymbol{\beta}_0||_2^2,$$

for any $\boldsymbol{\beta} \in \mathbb{B}$, some positive constant c_1 and all n sufficiently large.

(c) $G(t|\boldsymbol{\beta})$ and $G^{(1)}(t|\boldsymbol{\beta})$ satisfy the local Lipschitz conditions:

$$\sup_{|t|\leqslant c_0\sqrt{s\log(p\vee n)}}\left[G(t|\boldsymbol{\beta}_1) - G(t|\boldsymbol{\beta}_2)\right]^2 \leqslant c_1 s\log(p\vee n)||\boldsymbol{\beta}_1-\boldsymbol{\beta}_2||_2,$$

$$\sup_{|t|\leqslant c_0\sqrt{s\log(p\vee n)}}\left[G^{(1)}(t|\boldsymbol{\beta}_1) - G^{(1)}(t|\boldsymbol{\beta}_2)\right]^2 \leqslant c_1 s\log(p\vee n)||\boldsymbol{\beta}_1-\boldsymbol{\beta}_2||_2,$$

for any $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{B}$, for some positive constants c_0, c_1, and all n sufficiently large.

Remark (a). Our assumptions on the covariates in (A1) are similar to those in the literature on high-dimensional inference with random designs (e.g., Van de Geer et al. [2014], Belloni et al. [2015], among others). We assume $p \geqslant 2$. If $p = 1$, then the index model degenerates to a nonparametric model. The conditions in (A3) are standard assumptions on the kernel function for nonparametric smoothing. The assumptions in (A4) on the distributions of $\boldsymbol{x}^T\boldsymbol{\beta}$ are common for index models. Assumption (A2) involves restricted eigenvalue types assumptions and conditions on $\mathrm{E}(\boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\beta})$ and (A5) imposes conditions on the function $G(t|\boldsymbol{\beta})$ and $G^{(1)}(t|\boldsymbol{\beta})$. In Section S9 of the supplementary material, we verify that these key assumptions are satisfied when $\boldsymbol{x}$ follows the multivariate normal distribution in the high-dimensional setting.

Remark (b). Comparing with low-dimensional single-index models, conditions (A2) and (A5) are worthy of some discussions. The conditional expectations in both conditions depend on $\boldsymbol{x}^T\boldsymbol{\beta}$ and $\boldsymbol{\beta}$, possibly in a nonlinear fashion. As an example, in the multivariate normal distribution setting, the linearity condition $\mathrm{E}(\boldsymbol{x}^T\boldsymbol{\eta}|\boldsymbol{x}^T\boldsymbol{\beta}) = c_{\boldsymbol{\eta},\boldsymbol{\beta}}\boldsymbol{x}^T\boldsymbol{\beta}$, where $c_{\boldsymbol{\eta},\boldsymbol{\beta}}$ is a non-stochastic constant, plays an important role in the low-dimensional theory. In the high-dimensional setting, $c_{\boldsymbol{\eta},\boldsymbol{\beta}}$ (depending on $\boldsymbol{\beta}$, $\boldsymbol{\eta}$ nonlinearly) requires more careful analysis.

Remark (c). For identifiability, we assume that there exists a covariate with a nonzero coefficient. In practice, domain experts may help suggest such a candidate continuous covariate and the statisticians can run confirmatory analysis (e.g., comparing the conditional treatment effect conditional on this covariate) to verify if this is a viable choice. In the literature, another popular condition for identifiability is $||\boldsymbol{\beta}_0||_2 = 1$. However, it was also recognized (Yu and Ruppert [2002], Zhu and Xue [2006], Wang et al. [2010], among others) that technical derivation under this identifiability condition is more involved due to the fact $\boldsymbol{\beta}_0$ is a boundary point of a unit sphere and the derivative does not exist at $\boldsymbol{\beta}_0$. To handle this, the aforementioned literature suggested a delete-one-component approach. It was assumed that the true $\boldsymbol{\beta}$ has a positive component β_r. Let $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ and $\boldsymbol{\beta}^{(r)}$ be a $(p-1)$-subvector of $\boldsymbol{\beta}$ that excludes the r^{th} entry. Then we can write $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_{r-1}, \sqrt{1 - ||\boldsymbol{\beta}^{(r)}||_2^2}, \beta_{r+1}, \cdots, \beta_p)^T$. Thus the model can be reparametrized using the $(p-1)$-dimensional parameter $\boldsymbol{\beta}^{(r)}$. Under the assumption $||\boldsymbol{\beta}^{(r)}||_2 < 1$ (reasonable under the assumption that the underlying model has dimension at least two, otherwise it degenerates to a fully nonparametric model), the Jacobian matrix of the transformation can be computed as $\boldsymbol{J}_{\boldsymbol{\beta}^{(r)}} = (\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_p)^T$, where $\boldsymbol{\gamma}_s = \boldsymbol{e}_s$ (s^{th} column of the identity matrix $\boldsymbol{I}_p$), for $1 \leqslant s \leqslant p, s \neq r$, and $\boldsymbol{\gamma}_r = (1 - ||\boldsymbol{\beta}^{(r)}||_2^2)^{-1/2}(\beta_1, \cdots, \beta_{r-1}, -\sqrt{1 - ||\boldsymbol{\beta}^{(r)}||_2^2}, \beta_{r+1}, \cdots, \beta_p)^T$. Note that this transformation analysis also relies on knowing a covariate with a positive coefficient.

S3 Some useful definitions and lemmas

In this section, we introduce several useful definitions and lemmas which will be used in the proof of the theory. The proofs of these lemmas are given in Section S6.

Definition 1. *A random vector $\boldsymbol{x} \in \mathbb{R}^p$ is said to be sub-Gaussian with variance proxy σ_x^2 if* $\mathrm{E}\boldsymbol{x} = \mathbf{0}$*, and for each (fixed) unit vector $\boldsymbol{v} \in \mathbb{R}^p$,*

$$\mathrm{E}[\exp(s\boldsymbol{x}^T\boldsymbol{v})] \leqslant \exp\Big(\frac{\sigma_x^2 s^2}{2}\Big), \qquad \forall\ s \in \mathbb{R}.$$

An equivalent definition is that for each (fixed) unit vector $\boldsymbol{v} \in \mathbb{R}^p$ and any $t > 0$, $P(|\boldsymbol{x}^T\boldsymbol{v}| \geqslant t) \leqslant 2\exp\big(-\frac{t^2}{2\sigma_x^2}\big)$.

Property: Let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \in \mathbb{R}^p$ be independent sub-Gaussian random vectors in $\mathbb{R}^p$ with variance proxy σ_x^2. Then $\forall\ t > 0$, $P\Big(\max_{1\leqslant i\leqslant n} ||\boldsymbol{x}_i||_\infty > t\Big) \leqslant 2np\exp\big(-\frac{t^2}{2\sigma_x^2}\big)$. As a result,

$$P\Big(\max_{1\leqslant i\leqslant n} ||\boldsymbol{x}_i||_\infty > 2\sigma_x\sqrt{\log(np)}\Big) \leqslant 2\exp\big[-\log(np)\big]. \tag{S1}$$

Definition 2. *Let $L_k(\mathbb{P}_n) = \Big|n^{-1}\sum_{i=1}^n \boldsymbol{\gamma}^k(Z_i)\Big|^{1/k}$, $k = 1, 2$. For $\delta > 0$, the $\delta-$covering number $N(\delta, \Gamma, L_k(\mathbb{P}_n))$ of the class of functions Γ is the minimum number of $L_k(\mathbb{P}_n)-$balls with radius δ to cover Γ. The entropy is $H(\cdot, \Gamma, L_k(\mathbb{P}_n)) \triangleq \log\big[N(\delta, \Gamma, L_k(\mathbb{P}_n))\big]$.*

Definition 3. *A Rademacher sequence is a sequence $\epsilon_1, \cdots, \epsilon_n$ of i.i.d copies of a random variable ϵ taking values in $\{1, -1\}$, with $P(\epsilon = 1) = P(\epsilon = -1) = 1/2$.*

Lemma A1. *Under the assumptions of Theorem 1, there exist universal positive constants d_0 and d_1 such that $\big|\big|\boldsymbol{S}_n(\boldsymbol{\beta}_0, \widehat{G}, \widehat{\mathrm{E}})\big|\big|_\infty \leqslant d_0\sqrt{\frac{\log(p\vee n)}{n}}$ with probability at least $1 - \exp[-d_1\log(p \vee n)]$.*

Lemma A2. *If $\boldsymbol{x} \in \mathbb{R}^p$ is sub-Gaussian with variance proxy σ_x^2, then for any $\boldsymbol{\beta} \in \mathbb{B}$, $\mathrm{E}(\boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\beta})$ and $\boldsymbol{x} - \mathrm{E}(\boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\beta})$ are both sub-Gaussian, with variance proxy σ_x^2 and $2\sigma_x^2$, respectively. Furthermore, under assumption (A1), $\widetilde{\epsilon} = 2(2A-1)\big[\epsilon + g(\boldsymbol{x})\big]$ is sub-Gaussian* *with variance proxy* $4(\sigma_\epsilon^2 + M^2)$. *In addition, if* x *is a random variable such that* $|x| \leqslant \sigma_x$, *for some positive constant* σ_x, *and* y *is a sub-Gaussian random variable with variance proxy* σ_y^2, *then* $xy - \mathrm{E}(xy)$ *is sub-Gaussian with variance proxy no larger than* $4\sigma_x\sigma_y$.

Lemma A3. *Define the following four events:*

$$\mathcal{G}_n = \left\{ \max_{2 \leqslant j \leqslant p} \sup_{\boldsymbol{\beta} \in \mathbb{B}} n^{-1} \sum_{i=1}^n \left| \left[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1} | \boldsymbol{x}_i^T \boldsymbol{\beta}) \right]^T \boldsymbol{\theta}_j \right|^2 \leqslant d_0 \xi_2^{-2} \sigma_x^2 \right\},$$
$$\mathcal{H}_n = \left\{ \max_{2 \leqslant j \leqslant p} n^{-1} \sup_{\boldsymbol{\beta} \in \mathbb{B}} \sum_{i=1}^n \left| 2[\epsilon_i + g(\boldsymbol{x}_i)] * \left[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1} | \boldsymbol{x}_i^T \boldsymbol{\beta}) \right]^T \boldsymbol{\theta}_j \right|^2 \leqslant d_1 \xi_2^{-2} \sigma_x^2 (\sigma_\epsilon^2 + M^2) \right\},$$
$$\mathcal{J}_n = \left\{ \max_{1 \leqslant i \leqslant n} \sup_{\boldsymbol{\beta} \in \mathbb{B}_1} \left| \boldsymbol{x}_i^T \boldsymbol{\beta} \right| \leqslant 2 ||\boldsymbol{\beta}_0||_2 \sigma_x \sqrt{\log(p \vee n)} \right\},$$
$$\mathcal{K}_n = \left\{ \sup_{\boldsymbol{v} \in \mathbb{K}(p, s_0)} n^{-1} \sum_{i=1}^n \left| \boldsymbol{x}_i^T \boldsymbol{v} \right|^2 \leqslant 2\sigma_x^2 \right\},$$

for some positive constants $d_0 > 4$ *and* $d_1 > 256\sqrt{2}$. *Under the assumptions of Theorem 1, there exists some universal positive constant* c *such that* $P(\mathcal{G}_n \cap \mathcal{H}_n \cap \mathcal{J}_n \cap \mathcal{K}_n) \geqslant 1 - \exp[-c \log(p \vee n)]$, *for all* n *sufficiently large.*

Lemma A4. *Under the assumptions of Theorem 1, for any* $\boldsymbol{\beta} \in \mathbb{B}$,

$$\begin{aligned} G(\boldsymbol{x}^T\boldsymbol{\beta}|\boldsymbol{\beta}) - G(\boldsymbol{x}^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) =& f_0'(\boldsymbol{x}^T\boldsymbol{\beta}) \left[\boldsymbol{x}_{-1}^T \boldsymbol{\gamma} - \mathrm{E}(\boldsymbol{x}_{-1}^T \boldsymbol{\gamma} | \boldsymbol{x}^T \boldsymbol{\beta}) \right] \\ & - \left\{ h(\boldsymbol{x}_{-1}^T \boldsymbol{\gamma}) - \mathrm{E}\left[h(\boldsymbol{x}_{-1}^T \boldsymbol{\gamma}) | \boldsymbol{x}^T \boldsymbol{\beta} \right] \right\}, \end{aligned} \tag{S2}$$
$$\begin{aligned} G^{(1)}(\boldsymbol{x}^T\boldsymbol{\beta}|\boldsymbol{\beta}) - G^{(1)}(\boldsymbol{x}^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) =& f_0''(\boldsymbol{x}^T\boldsymbol{\beta}) \left[\boldsymbol{x}_{-1}^T \boldsymbol{\gamma} - \mathrm{E}(\boldsymbol{x}_{-1}^T \boldsymbol{\gamma} | \boldsymbol{x}^T \boldsymbol{\beta}) \right] - h_1(\boldsymbol{x}_{-1}^T \boldsymbol{\gamma}) \\ & - f_0'(\boldsymbol{x}^T\boldsymbol{\beta}) \mathrm{E}^{(1)}(\boldsymbol{x}_{-1}^T \boldsymbol{\gamma} | \boldsymbol{x}^T \boldsymbol{\beta}) + \mathrm{E}^{(1)}\left[h(\boldsymbol{x}_{-1}^T \boldsymbol{\gamma}) | \boldsymbol{x}^T \boldsymbol{\beta} \right], \end{aligned} \tag{S3}$$

where $\boldsymbol{\gamma} = \boldsymbol{\beta}_{-1} - \boldsymbol{\beta}_{0,-1}$, $h(u) = \int_0^u a f_0''(a + \boldsymbol{x}^T\boldsymbol{\beta}_0) da$, $h_1(u) = \int_0^u a f_0'''(a + \boldsymbol{x}^T\boldsymbol{\beta}_0) da$, *and* $\mathrm{E}^{(1)}(\cdot | \boldsymbol{x}^T\boldsymbol{\beta} = t)$ *is the first derivative of* $\mathrm{E}(\cdot | \boldsymbol{x}^T\boldsymbol{\beta} = t)$ *with respect to* t. *Moreover, there exist* *universal positive constants c_1 and c_2 such that for all n sufficiently large,*

$$\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2\in\mathbb{B}}\left[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_1|\boldsymbol{\beta}_1)-G(\boldsymbol{x}_i^T\boldsymbol{\beta}_2|\boldsymbol{\beta}_2)\right]^2\leqslant c_1||\boldsymbol{\beta}_1-\boldsymbol{\beta}_2||_2 s\log(p\vee n), \tag{S4}$$

$$\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2\in\mathbb{B}}\left[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_1|\boldsymbol{\beta}_1)-G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_2|\boldsymbol{\beta}_2)\right]^2\leqslant c_1||\boldsymbol{\beta}_1-\boldsymbol{\beta}_2||_2 s\log(p\vee n), \tag{S5}$$

$$n^{-1}\sum_{i=1}^n\left[G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right]^2\leqslant c_1||\boldsymbol{\beta}-\boldsymbol{\beta}_0||_2^2,\ \forall\ \boldsymbol{\beta}\in\mathbb{B} \tag{S6}$$

with probability at least $1-\exp[-c_2\log(p\vee n)]$.

Lemma A5. *Under the assumptions of Theorem 1, there exist universal positive constants c_0 and c_1 such that for all n sufficiently large,*

$$P\Big(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}}\big|\widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big|\geqslant c_0h^2\Big)\leqslant\exp[-c_1\log(p\vee n)].$$

Lemma A6. *Under the assumptions of Theorem 1, there exist universal positive constants c_0 and c_1 such that for all n sufficiently large,*

$$P\Big(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}}\big|\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big|\geqslant c_0h\Big)\leqslant\exp[-c_1\log(p\vee n)].$$

Lemma A7. *Under the assumptions of Theorem 1, there exist universal positive constants c_0, c_1 such that for all n sufficiently large,*

$$P\Big(\max_{1\leqslant i\leqslant n}\sup_{\substack{\boldsymbol{\beta}\in\mathbb{B}\\ \boldsymbol{v}\in\mathbb{K}(p,2ks)}}\Big|\Big[\widehat{\mathrm{E}}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})\Big]^T\boldsymbol{v}\Big|\geqslant c_0h^2\Big)\leqslant\exp[-c_1\log(p\vee n)],$$

$$P\Big(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}}\Big|\Big|\widehat{\mathrm{E}}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})\Big|\Big|_\infty\geqslant c_0h^2\Big)\leqslant\exp[-c_1\log(p\vee n)].$$

Furthermore, if $\widetilde{s}\log p\leqslant d_0n$ for some positive constant d_0, where $s_j=||\boldsymbol{d}_{0j}||_0$, $\widetilde{s}=\max_{2\leqslant j\leqslant p}s_j$, then there exist universal positive constants d_1, d_2 such that for all n suffi-

ciently large,

$$P\Big(\max_{1\leqslant i\leqslant n}\sup_{\substack{\boldsymbol{\beta}\in\mathbb{B}\\ \boldsymbol{v}\in\mathbb{K}(p,2ks+\widetilde{s})}}\Big|\Big[\widehat{\mathrm{E}}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})\Big]^T\boldsymbol{v}\Big|\geqslant d_1h^2\Big)\leqslant\exp[-d_2\log(p\vee n)].$$

Lemma A8. *Under the assumptions of Theorem 2, there exist universal positive constants c_0, c_1, such that for all n sufficiently large,*

$$P\left(\max_{2\leqslant j\leqslant p}\left|n^{-1/2}\sum_{i=1}^n\boldsymbol{\theta}_j^T\boldsymbol{\gamma}(Z_i,\widehat{\boldsymbol{\beta}},\widehat{G}^{(1)})\right|\geqslant c_0\left[h^2\log(p\vee n)\right]^{1/4}\right)\leqslant\exp(-c_1\log p),\tag{S7}$$

$$P\left(\left\|n^{-1/2}\sum_{i=1}^n\boldsymbol{\gamma}(Z_i,\widehat{\boldsymbol{\beta}},\widehat{G}^{(1)})\right\|_\infty\geqslant c_0\left[h^2\log(p\vee n)\right]^{1/4}\right)\leqslant\exp(-c_1\log p),\tag{S8}$$

where $Z_i=(\boldsymbol{x}_i,\epsilon_i,A_i)$, $\boldsymbol{\gamma}(Z_i,\widehat{\boldsymbol{\beta}},\widehat{G}^{(1)})=\big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\big]\widetilde{\epsilon}_i\big[\boldsymbol{x}_{i,-1}-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]$, with $\widetilde{\epsilon}_i=2(2A_i-1)[\epsilon_i+g(\boldsymbol{x}_i)]$.

Lemma A9. *Under the assumptions of Theorem 2, there exist universal positive constants c_0, c_1 such that for all n sufficiently large,*

$$P\left(\max_{2\leqslant j\leqslant p}\left|n^{-1/2}\sum_{i=1}^n\boldsymbol{\theta}_j^T\boldsymbol{\nu}_1(Z_i,\widehat{\boldsymbol{\beta}},\widehat{G}^{(1)})\right|\geqslant c_0sh^3\sqrt{n}\right)\leqslant\exp(-c_1\log p),\tag{S9}$$

$$P\left(\left\|n^{-1/2}\sum_{i=1}^n\boldsymbol{\nu}_1(Z_i,\widehat{\boldsymbol{\beta}},\widehat{G}^{(1)})\right\|_\infty\geqslant c_0sh^3\sqrt{n}\right)\leqslant\exp(-c_1\log p),\tag{S10}$$

$$P\left(\max_{2\leqslant j\leqslant p}\left|n^{-1/2}\sum_{i=1}^n\boldsymbol{\theta}_j^T\boldsymbol{\nu}_2(Z_i,\widehat{\boldsymbol{\beta}},\widehat{G},\widehat{G}^{(1)})\right|\geqslant c_0sh^3\sqrt{n}\right)\leqslant\exp(-c_1\log p),\tag{S11}$$

$$P\left(\left\|n^{-1/2}\sum_{i=1}^n\boldsymbol{\nu}_2(Z_i,\widehat{\boldsymbol{\beta}},\widehat{G},\widehat{G}^{(1)})\right\|_\infty\geqslant c_0sh^3\sqrt{n}\right)\leqslant\exp(-c_1\log p),\tag{S12}$$

where $Z_i=(\boldsymbol{x}_i,\epsilon_i,A_i)$, $\boldsymbol{\nu}_1(Z_i,\widehat{\boldsymbol{\beta}},\widehat{G}^{(1)})=\big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)-G(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})-G^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{x}}_{i,-1}^T(\boldsymbol{\beta}_{0,-1}-\widehat{\boldsymbol{\beta}}_{-1})\big]\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{x}}_{i,-1}$, $\boldsymbol{\nu}_2(Z_i,\widehat{\boldsymbol{\beta}},\widehat{G},\widehat{G}^{(1)})=\big[G(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})-\widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\big]\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{x}}_{i,-1}$, with $\widehat{\boldsymbol{x}}_{i,-1}=\boldsymbol{x}_{i,-1}-\widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})$.

Lemma A10. *Under the assumptions of Theorem 1, there exist universal positive constants*

c_0 *and* c_1 *such that for all* n *sufficiently large,*

$$P\left(\max_{2\leqslant j\leqslant p}\left|n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\theta}_j^T\boldsymbol{\xi}(Z_i,\widehat{\boldsymbol{\beta}},\widehat{\mathrm{E}})\right|\geqslant c_0h\sqrt{s\log(p\vee n)}\right)\leqslant\exp(-c_1\log p), \tag{S13}$$

$$P\left(\left\|n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\xi}(Z_i,\boldsymbol{\beta},\widehat{\mathrm{E}})\right\|_\infty\geqslant c_0h\sqrt{s\log(p\vee n)}\right)\leqslant\exp(-c_1\log p), \tag{S14}$$

where $Z_i=(\boldsymbol{x}_i,\epsilon_i,A_i)$, $\boldsymbol{\xi}(Z_i,\widehat{\boldsymbol{\beta}},\widehat{\mathrm{E}})=\widetilde{\epsilon}_iG^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\left[\widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\right]$, *with* $\widetilde{\epsilon}_i=2(2A_i-1)[\epsilon_i+g(\boldsymbol{x}_i)]$.

Lemma A11. *Under the assumptions of Lemma 2, there exist universal positive constants* c_0 *and* c_1 *such that for all* n *sufficiently large,*

$$P\Big(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\big|\widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\big|\geqslant c_0sh^2\sqrt{\log(p\vee n)}\Big)\leqslant\exp(-c_1\log p), \tag{S15}$$

$$P\Big(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\big|\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\big|\geqslant c_0h\Big)\leqslant\exp(-c_1\log p), \tag{S16}$$

$$P\Big(\max_{1\leqslant i\leqslant n}\sup_{\substack{\boldsymbol{\beta}\in\mathbb{B}\\ \boldsymbol{v}\in\mathbb{K}(p,2ks+\widetilde{s})}}\big|[\widehat{\mathrm{E}}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^T\boldsymbol{v}\big|\geqslant c_0sh^2\sqrt{\log(p\vee n)}\Big)\leqslant\exp(-c_1\log p). \tag{S17}$$

Lemma A12. *Under the assumptions of Lemma 2, there exist universal positive constants* c_0, c_1, *such that all* n *sufficiently large,*

$$P\Big(\sup_{\substack{\boldsymbol{\beta}\in\mathbb{B}\\ \boldsymbol{v}\in\mathbb{K}(p,2ks+\widetilde{s})}}\Big|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{v}^T\big[\widehat{\mathrm{E}}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]\big[\widehat{\mathrm{E}}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{v}\Big| \geqslant c_0sh^4\Big)\leqslant\exp(-c_1\log p), \tag{S18}$$

$$P\Big(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\Big\|\frac{1}{n}\sum_{i=1}^{n}\big[\widehat{\mathrm{E}}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]\big[\widehat{\mathrm{E}}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\Big\|_\infty \geqslant c_0sh^4\Big)\leqslant\exp(-c_1\log p). \tag{S19}$$

Lemma A13 below gives an alternative expression for $\boldsymbol{\phi}_{0j}$.

Lemma A13. *Under assumption (A2), we have*

$$\boldsymbol{\phi}_{0j} = \Big(- (\boldsymbol{d}_{0j})^T_{1:(j-2)}, 1, -(\boldsymbol{d}_{0j})^T_{(j-1):(p-1)} \Big)^T,$$

and $||\boldsymbol{\theta}_j||_0 \leqslant \widetilde{s}+1$ *uniformly in* $j = 2, \ldots, p$. *Under Assumption (A1) and Assumption (A2), we have* $\tau_{0j}^{-2} \leqslant ||\boldsymbol{\theta}_j||_2 \leqslant \xi_2^{-1}$ *and* $\tau_{0j}^2 \leqslant b^2 \xi_1$ *uniformly in* $j = 2, \ldots, p$.

Lemma A14. *Under the assumptions of Theorem 2,* $||\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\Theta}^T \boldsymbol{\Lambda} \boldsymbol{\Theta}||_\infty = o_p(1)$.

S4 Proofs of results in Section 3.1 of the main paper

Proof of Lemma 1. Note that

$$\begin{aligned}
\boldsymbol{S}_n(\boldsymbol{\beta}, \widehat{G}, \widehat{\mathrm{E}}) &= -\frac{1}{n}\sum_{i=1}^n \big[\widetilde{Y}_i - \widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big[\boldsymbol{x}_{i,-1} - \widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})\big] \\
&= -\frac{1}{n}\sum_{i=1}^n \widetilde{\epsilon}_i \widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big[\boldsymbol{x}_{i,-1} - \widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})\big] \\
&\quad -\frac{1}{n}\sum_{i=1}^n \big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) - \widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big[\boldsymbol{x}_{i,-1} - \widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})\big],
\end{aligned}$$

where $\widetilde{\epsilon}_i = 2(2A_i - 1)[\epsilon_i + g(\boldsymbol{x}_i)]$. Denote $\boldsymbol{\gamma} = \boldsymbol{\beta}_{-1} - \boldsymbol{\beta}_{0,-1}$. We have

$$
\begin{aligned}
&\langle \boldsymbol{S}_n(\boldsymbol{\beta}, \widehat{G}, \widehat{\mathrm{E}}) - \boldsymbol{S}_n(\boldsymbol{\beta}_0, \widehat{G}, \widehat{\mathrm{E}}), \boldsymbol{\gamma}\rangle \\
=& -\frac{1}{n}\sum_{i=1}^n \widetilde{\epsilon}_i \widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big[\widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - \widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})\big]^T\boldsymbol{\gamma} \\
& -\frac{1}{n}\sum_{i=1}^n \widetilde{\epsilon}_i \big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - \widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\big]\big[\boldsymbol{x}_{i,-1} - \widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\gamma} \\
& -\frac{1}{n}\sum_{i=1}^n \big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) - \widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big[\widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - \widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})\big]^T\boldsymbol{\gamma} \\
& -\frac{1}{n}\sum_{i=1}^n \big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) - \widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]\big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - \widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\big]\big[\boldsymbol{x}_{i,-1} - \widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\gamma} \\
& -\frac{1}{n}\sum_{i=1}^n \big[\widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) - \widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\big[\boldsymbol{x}_{i,-1} - \widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\gamma} \\
\triangleq& -\sum_{k=1}^5 A_k(\boldsymbol{\beta}),
\end{aligned}
$$

where the definition of A_k's, $k = 1, \cdots, 5$, is clear from the context. Each A_k can be further decomposed.

$$
\begin{aligned}
& A_1(\boldsymbol{\beta}) \\
= & \frac{1}{n}\sum_{i=1}^n \widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big[\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})\big]^T\boldsymbol{\gamma} \\
& +\frac{1}{n}\sum_{i=1}^n \widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big[\widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - \widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0) + \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})\big]^T\boldsymbol{\gamma} \\
& +\frac{1}{n}\sum_{i=1}^n \widetilde{\epsilon}_i \big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]\big[\widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - \widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})\big]^T\boldsymbol{\gamma} \\
\triangleq & \sum_{l=1}^3 A_{1l}(\boldsymbol{\beta}).
\end{aligned}
$$

$$
\begin{aligned}
& A_2(\boldsymbol{\beta}) \\
= \;& \frac{1}{n}\sum_{i=1}^{n}\widetilde{\epsilon}_i\big[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\big]\big[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\gamma} \\
& + \frac{1}{n}\sum_{i=1}^{n}\widetilde{\epsilon}_i\big[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\big]\big[\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - \widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\gamma} \\
& + \frac{1}{n}\sum_{i=1}^{n}\widetilde{\epsilon}_i\big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - \widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) + G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\big] \\
& \qquad\qquad * \big[\boldsymbol{x}_{i,-1} - \widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\gamma} \\
\triangleq \;& \sum_{l=1}^{3} A_{2l}(\boldsymbol{\beta}),
\end{aligned}
$$

$$
\begin{aligned}
& A_3(\boldsymbol{\beta}) \\
= \;& \frac{1}{n}\sum_{i=1}^{n}\big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) - G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big[\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})\big]^T\boldsymbol{\gamma} \\
& + \frac{1}{n}\sum_{i=1}^{n}\big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) - G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) \\
& \qquad * \big[\widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - \widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0) + \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})\big]^T\boldsymbol{\gamma} \\
& + \frac{1}{n}\sum_{i=1}^{n}\big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) - G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]\big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big] \\
& \qquad\qquad * \big[\widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - \widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})\big]^T\boldsymbol{\gamma} \\
& + \frac{1}{n}\sum_{i=1}^{n}\big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - \widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big[\widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - \widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})\big]^T\boldsymbol{\gamma} \\
\triangleq \;& \sum_{l=1}^{4} A_{3l}(\boldsymbol{\beta}).
\end{aligned}
$$

$$
\begin{aligned}
& A_4(\boldsymbol{\beta}) + A_5(\boldsymbol{\beta}) \\
= \; & \frac{1}{n}\sum_{i=1}^{n}\big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) - G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\gamma} \\
& + \frac{1}{n}\sum_{i=1}^{n}\big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) - G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big[\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - \widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\gamma} \\
& + \frac{1}{n}\sum_{i=1}^{n}\big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) - G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]\big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]\big[\boldsymbol{x}_{i,-1} - \widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\gamma} \\
& + \frac{1}{n}\sum_{i=1}^{n}\big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - \widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]\big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - \widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\big]\big[\boldsymbol{x}_{i,-1} - \widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\gamma} \\
& + \frac{1}{n}\sum_{i=1}^{n}\big[\widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) - G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) - \widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) + G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big] \\
& \qquad * \widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\big[\boldsymbol{x}_{i,-1} - \widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\gamma} \\
\triangleq \; & \sum_{l=1}^{5} B_l(\boldsymbol{\beta}).
\end{aligned}
$$

The proof involves evaluating the order of A_k $(k = 1, 2, 3)$ and $A_4 + A_5$. We provide the details of analyzing $B_1(\boldsymbol{\beta})$ and $B_2(\boldsymbol{\beta})$ which are two of the most challenging terms to study. All the other terms can be handled similarly. First, for the function $h(\cdot)$ defined in Lemma A4, we have $h(\boldsymbol{x}_{-1}^T\boldsymbol{\gamma}) = \int_0^{\boldsymbol{x}_{-1}^T\boldsymbol{\gamma}} a f_0''(a + \boldsymbol{x}^T\boldsymbol{\beta}_0)da = \frac{1}{2}f_0''(\boldsymbol{x}^T\boldsymbol{\beta}_1)(\boldsymbol{x}_{-1}^T\boldsymbol{\gamma})^2$ for some $\boldsymbol{\beta}_1$ between $\boldsymbol{x}^T\boldsymbol{\beta}_0$ and $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}^T\boldsymbol{\beta}_0 + \boldsymbol{x}_{-1}^T\boldsymbol{\gamma}$. By (S2) in Lemma A4, we can write $-B_1 =$

$\sum_{q=1}^{6} B_{1q}$, where

$$B_{11} = n^{-1}\sum_{i=1}^{n}\big[f_0'(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^2\boldsymbol{\gamma}^T\big[\boldsymbol{x}_i - \mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]\big[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\gamma},$$
$$B_{12} = n^{-1}\sum_{i=1}^{n} f_0'(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)[f_0'(\boldsymbol{x}_i^T\boldsymbol{\beta}) - f_0'(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]\boldsymbol{\gamma}^T\big[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]\big[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\gamma},$$
$$B_{13} = n^{-1}\sum_{i=1}^{n}[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]f_0'(\boldsymbol{x}_i^T\boldsymbol{\beta})\Big\{\big[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\gamma}\Big\}^2,$$
$$B_{14} = n^{-1}\sum_{i=1}^{n} G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})f_0'(\boldsymbol{x}_i^T\boldsymbol{\beta})\boldsymbol{\gamma}^T\big[\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})\big]\big[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_{i,-1}^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\gamma},$$
$$B_{15} = (2n)^{-1}\sum_{i=1}^{n} G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})f_0''(\boldsymbol{x}_i^T\boldsymbol{\beta}_2)(\boldsymbol{x}_{i,-1}^T\boldsymbol{\gamma})^2\big[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\gamma},$$
$$B_{16} = (2n)^{-1}\sum_{i=1}^{n} G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\mathrm{E}\big[f_0''(\boldsymbol{x}_i^T\boldsymbol{\beta}_1)(\boldsymbol{x}_{i,-1}^T\boldsymbol{\gamma})^2|\boldsymbol{x}_i^T\boldsymbol{\beta}\big]\big[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\gamma},$$

for some $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$. Note that $G(\boldsymbol{x}^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) = f(\boldsymbol{x}^T\boldsymbol{\beta}_0)$, and $G^{(1)}(\boldsymbol{x}^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) = f'(\boldsymbol{x}^T\boldsymbol{\beta}_0)$.

Observe that

$$\begin{aligned}
&B_{11}\\
=&n^{-1}\sum_{i=1}^{n}\big[f_0'(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^2\boldsymbol{\gamma}^T\mathrm{E}\big[\mathrm{Cov}(\boldsymbol{x}_{-1}|\boldsymbol{x}^T\boldsymbol{\beta}_0)\big]\boldsymbol{\gamma}\\
&+n^{-1}\sum_{i=1}^{n}\big[f_0'(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^2\boldsymbol{\gamma}^T\big\{\big[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]\big[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T - \mathrm{E}\big[\mathrm{Cov}(\boldsymbol{x}_{-1}|\boldsymbol{x}^T\boldsymbol{\beta}_0)\big]\big\}\boldsymbol{\gamma}\\
=&B_{111} + B_{112},
\end{aligned}$$

where the definitions of B_{111} and B_{112} are clear from the context. Assumption (A1)-(b) implies that $P\big(n^{-1}\sum_{i=1}^{n}[f_0'(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2 \leqslant a^2/2\big) \leqslant \exp\left(\frac{-na^2}{4b^2}\right)$, according to Hoeffding's inequality. Combined with Assumption (A2), we have $B_{111} \geqslant a^2\xi_0||\boldsymbol{\gamma}||_2^2/2$, with probability at least $1-\exp\left(\frac{-na^2}{4b^2}\right)$. Note that $\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)$ is sub-Gaussian by Lemma A2. Lemma B2 implies $|B_{112}| \leqslant c_0||\boldsymbol{\gamma}||_2^2\sqrt{\frac{s\log p}{n}}$, with probability at least $1-\exp(-c_1 s\log p)$, for some positive constants c_0, c_1, and all n sufficiently large. Therefore, we obtain that $B_{11} \geqslant ||\boldsymbol{\gamma}||_2^2 \left(a^2\xi_0/2 - \sqrt{\frac{s\log p}{n}}\right)$ with probability at least $1-\exp(-c_1 s\log p)$, for some positive constant c_1, and all n sufficiently large.

To evaluate B_{12}, we observe that there exists a point $\boldsymbol{\beta}^r$ between $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}$ such that

$$\begin{aligned}
|B_{12}| =& \Big| n^{-1}\sum_{i=1}^n f_0'(\boldsymbol{x}_i^T\boldsymbol{\beta}_0) f_0''(\boldsymbol{x}_i^T\boldsymbol{\beta}^r)(\boldsymbol{x}_{i,-1}^T\boldsymbol{\gamma})\boldsymbol{\gamma}^T\big[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]\big[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\gamma}\Big| \\
\leqslant& C n^{-1}\sum_{i=1}^n |\boldsymbol{x}_{i,-1}^T\boldsymbol{\gamma}| * \boldsymbol{\gamma}^T\big[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]\big[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\gamma} \\
\leqslant& C\sqrt{n^{-1}\sum_{i=1}^n(\boldsymbol{x}_{i,-1}^T\boldsymbol{\gamma})^2} * \sqrt{n^{-1}\sum_{i=1}^n\Big\{\big[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\gamma}\Big\}^4},
\end{aligned}$$

for some positive constant C, given Assumption (A1)-(b). Recall that $\widetilde{\boldsymbol{x}}_{i,-1} = \boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)$. Lemma B3 indicates that

$$P\left(\Big|n^{-1}\sum_{i=1}^n\big[|\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\gamma}|^4 - \mathrm{E}(|\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\gamma}|^4)\big]\Big| \geqslant c_1||\boldsymbol{\gamma}||_2^4\left[\sqrt{\frac{s\log p}{n}} \vee \frac{s^2\log^2 p}{n}\right]\right) \leqslant \exp(-c_2 s\log p),$$

for some positive constants c_1, c_2, and all n sufficiently large. Hence $|B_{12}| \leqslant c_1||\boldsymbol{\gamma}||_2 * ||\boldsymbol{\gamma}||_2^2 = c_1||\boldsymbol{\gamma}||_2^3$, with probability at least $1-\exp(-c_2 s\log p)$, for some positive constants c_1, c_2, and all n sufficiently large.

Hölder's Inequality and Assumption (A5)-(b) imply that

$$\begin{aligned}
|B_{13}| \leqslant& b\Big(n^{-1}\sum_{i=1}^n|G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)|^2\Big)^{1/2} * \Big(n^{-1}\sum_{i=1}^n\big|\big[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\gamma}\big|^4\Big)^{1/2} \\
\leqslant& c_1||\boldsymbol{\gamma}||_2 * \Big(n^{-1}\sum_{i=1}^n\Big|\big[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\gamma}\Big|^4\Big)^{1/2},
\end{aligned}$$

with probability at least $1-\exp(-c_2\log p)$, for some positive constants c_1, c_2, and all n sufficiently large. Similarly as the derivation for B_{12}, $|B_{13}| \leqslant c_1||\boldsymbol{\gamma}||_2 * ||\boldsymbol{\gamma}||_2^2 = c_1||\boldsymbol{\gamma}||_2^3$, with probability at least $1-\exp(-c_2\log p)$, for some positive constants c_1, c_2, and all n sufficiently large.

To evaluate B_{14}, we observe that

$$
\begin{aligned}
&|B_{14}| \\
\leqslant& \frac{b^2}{n}\sum_{i=1}^n \Big|\boldsymbol{\gamma}^T\big[\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})\big]\big[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\gamma}\Big| \\
\leqslant& \frac{b^2}{n}\sum_{i=1}^n C\big[|\boldsymbol{x}_{i,-1}^T\boldsymbol{\gamma}| + (|\boldsymbol{x}_i^T\boldsymbol{\beta}| + |\boldsymbol{x}_i^T\boldsymbol{\beta}_0|) * ||\boldsymbol{\gamma}||_2\big]\Big|\big[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\gamma}\Big| \\
\leqslant& b^2C||\boldsymbol{\gamma}||_2\sqrt{\frac{2}{n}\sum_{i=1}^n(\boldsymbol{x}_{i,-1}^T\boldsymbol{\gamma})^2 + (|\boldsymbol{x}_i^T\boldsymbol{\beta}|^2 + |\boldsymbol{x}_i^T\boldsymbol{\beta}_0|^2) * ||\boldsymbol{\gamma}||_2^2}\sqrt{\frac{1}{n}\sum_{i=1}^n\Big\{\big[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\gamma}\Big\}^2},
\end{aligned}
$$

where the second inequality applies Assumption (A2)-(c). Since $\boldsymbol{x}_{i,-1}$ is mean-zero sub-Gaussian with variance proxy σ_x^2, similarly as previous steps, we have

$$\frac{1}{n}\sum_{i=1}^n(\boldsymbol{x}_{i,-1}^T\boldsymbol{\gamma})^2 \leqslant \left(\xi_3 + c\sqrt{\frac{s\log p}{n}}\right)||\boldsymbol{\gamma}||_2^2,$$

with probability at least $1-\exp(-c_1 s\log p)$, for some positive constants c, c_1. Since $[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]$ is also sub-Gaussian by Lemma A2, similarly we have

$$\frac{1}{n}\sum_{i=1}^n\Big\{\big[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\gamma}\Big\}^2 \leqslant c_2||\boldsymbol{\gamma}||_2^2,$$

with probability at least $1-\exp(-c_1 s\log p)$ for some positive constants c_1, c_2, and all n sufficiently large. It follows that $|B_{14}| \leqslant c_0||\boldsymbol{\gamma}||_2^3$. Since $\boldsymbol{x}_{i,-1}$ and $[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]$ are both sub-Gaussian, we apply the same techniques to $|B_{15}|$. Lemma B3 ensures that $|B_{15}| \leqslant c_0\sigma_x^3||\boldsymbol{\gamma}||_2^3$ with probability at least $1-\exp(-c_1 s\log p)$, for some positive constants c_0, c_1, and all n sufficiently large.

To bound $|B_{16}|$, Assumption (A2)-(a), (A4) and (A5)-(a) imply that

$$\begin{aligned}|B_{16}| \leqslant & \frac{c}{n}\sum_{i=1}^{n}\left|\boldsymbol{\gamma}^T \mathrm{E}(\boldsymbol{x}_{i,-1}\boldsymbol{x}_{i,-1}^T|\boldsymbol{x}_i^T\boldsymbol{\beta})\boldsymbol{\gamma}\big[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\gamma}\right| \\ \leqslant & c\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|\boldsymbol{\gamma}^T \mathrm{E}(\boldsymbol{x}_{i,-1}\boldsymbol{x}_{i,-1}^T|\boldsymbol{x}_i^T\boldsymbol{\beta})\boldsymbol{\gamma}\right|^2} * \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\gamma}\right|^2} \\ \leqslant & c_0\sqrt{\xi_4}||\boldsymbol{\gamma}||_2^3,\end{aligned}$$

holds with probability at least $1 - \exp[-c_1\log(p \vee n)] - \exp(-c_1 s\log p)$ for some positive constants c_0, c_1, and all n sufficiently large. In the above, the last inequality applies the sub-Gaussian property of $[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]$, similarly as the derivation for B_{14}.

Combining all the preceding results, we conclude that $-B_1 \geqslant c_0\left(||\boldsymbol{\gamma}||_2^2 - ||\boldsymbol{\gamma}||_2^2\sqrt{\frac{\log p}{n}}\right)$ with probability at least $1 - \exp(-c_1\log p)$, and universal positive constants c_0 and c_1, for all n sufficiently large.

Assumption (A5)-(a), Lemma A4 and Lemma A7 imply that

$$\begin{aligned}|B_2| \leqslant & \, b\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}}\left|\big[\widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}) - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})\big]^T\boldsymbol{\gamma}\right| * n^{-1}\sum_{i=1}^{n}|G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) - G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})| \\ \leqslant & \, d_1h^2||\boldsymbol{\gamma}||_2^2,\end{aligned}$$

with probability at least $1-\exp[-c_2\log(p \vee n)]$, for some universal positive constants d_1, c_2, and all n sufficiently large. Similarly, we can show that for some positive universal constants

d_1 and c_1,

$$
\begin{aligned}
&|A_{11}| \leqslant d_1||\boldsymbol{\gamma}||_2^2\sqrt{\frac{\log p}{n}}, && |A_{12}| \leqslant d_1h^2||\boldsymbol{\gamma}||_2\sqrt{\frac{\log p}{n}}, && |A_{13}| \leqslant d_1h||\boldsymbol{\gamma}||_2^2\sqrt{\frac{\log p}{n}},\\
&|A_{21}| \leqslant d_1||\boldsymbol{\gamma}||_2^2\sqrt{\frac{\log p}{n}}, && |A_{22}| \leqslant d_1h^2||\boldsymbol{\gamma}||_2^2\sqrt{\frac{\log p}{n}}, && |A_{23}| \leqslant d_1h||\boldsymbol{\gamma}||_2\sqrt{\frac{\log p}{n}},\\
&|A_{31}| \leqslant d_1||\boldsymbol{\gamma}||_2^3, && |A_{32}| \leqslant d_1h^2||\boldsymbol{\gamma}||_2^2, && |A_{33}| \leqslant d_1h||\boldsymbol{\gamma}||_2^3,\\
&|A_{34}| \leqslant d_1h^2||\boldsymbol{\gamma}||_2^2, && |B_3| \leqslant d_1h||\boldsymbol{\gamma}||_2^2, && |B_4| \leqslant d_1h^3||\boldsymbol{\gamma}||_2,\\
&|B_5| \leqslant d_1h^2||\boldsymbol{\gamma}||_2,
\end{aligned}
$$

hold with probability at least $1-\exp(-c_1\log p)$, for all n sufficiently large. Since $n^{-1}\log p = O(h^5)$, there exist some universal positive constants c_0, c_1, c_2 and $r \leqslant 1$ such that

$$\left\langle \boldsymbol{S}_n(\boldsymbol{\beta}, \widehat{G}, \widehat{\mathrm{E}}) - \boldsymbol{S}_n(\boldsymbol{\beta}_0, \widehat{G}, \widehat{\mathrm{E}}), \boldsymbol{\gamma} \right\rangle \geqslant c_0||\boldsymbol{\gamma}||_2^2 - c_1h^2||\boldsymbol{\gamma}||_2,$$

with probability at least $1-\exp(-c_1\log p)$, for any $\boldsymbol{\beta} \in \mathbb{B}$ and all n sufficiently large. □

Proof of Theorem 1. By the definition of $\widehat{\boldsymbol{\beta}} = (1, \widehat{\boldsymbol{\beta}}_{-1}^T)^T$, we have

$$\left\langle \boldsymbol{S}_n(\widehat{\boldsymbol{\beta}}, \widehat{G}, \widehat{\mathrm{E}}) + \lambda\widehat{\boldsymbol{\kappa}}, \boldsymbol{\beta}_{-1} - \widehat{\boldsymbol{\beta}}_{-1} \right\rangle = 0, \tag{S20}$$

for all feasible $\boldsymbol{\beta}$, where $\widehat{\boldsymbol{\kappa}} \in \partial||\widehat{\boldsymbol{\beta}}_{-1}||_1$. In particular, $\left\langle \boldsymbol{S}_n(\widehat{\boldsymbol{\beta}}, \widehat{G}, \widehat{\mathrm{E}}) + \lambda\widehat{\boldsymbol{\kappa}}, \boldsymbol{\beta}_{0,-1} - \widehat{\boldsymbol{\beta}}_{-1} \right\rangle = 0$. By the property of convex function, we know that $||\boldsymbol{\beta}_0||_1 - ||\widehat{\boldsymbol{\beta}}||_1 \geqslant \left\langle \widehat{\boldsymbol{\kappa}}, \boldsymbol{\beta}_{0,-1} - \widehat{\boldsymbol{\beta}}_{-1} \right\rangle$, for any $\widehat{\boldsymbol{\kappa}} \in \partial||\widehat{\boldsymbol{\beta}}_{-1}||_1$. Combining this with (S20), we have

$$\left\langle \boldsymbol{S}_n(\widehat{\boldsymbol{\beta}}, \widehat{G}, \widehat{\mathrm{E}}), \widehat{\boldsymbol{\eta}} \right\rangle = \lambda\left\langle \widehat{\boldsymbol{\kappa}}, -\widehat{\boldsymbol{\eta}} \right\rangle \leqslant \lambda(||\boldsymbol{\beta}_{0,-1}||_1 - ||\widehat{\boldsymbol{\beta}}_{-1}||_1). \tag{S21}$$

where $\widehat{\boldsymbol{\eta}} = \widehat{\boldsymbol{\beta}}_{-1} - \boldsymbol{\beta}_{0,-1}$. Applying the local restricted strong convexity condition established in Lemma 1 to $\left\langle \boldsymbol{S}_n(\boldsymbol{\beta}, \widehat{G}, \widehat{\mathrm{E}}) - \boldsymbol{S}_n(\boldsymbol{\beta}_0, \widehat{G}, \widehat{\mathrm{E}}), \widehat{\boldsymbol{\eta}} \right\rangle$, we obtain

$$\begin{aligned} c_0||\widehat{\boldsymbol{\eta}}||_2^2 - c_1 h^2 ||\widehat{\boldsymbol{\eta}}||_2 &\leqslant \left\langle \boldsymbol{S}_n(\widehat{\boldsymbol{\beta}}, \widehat{G}, \widehat{\mathrm{E}}), \widehat{\boldsymbol{\eta}} \right\rangle - \left\langle \boldsymbol{S}_n(\boldsymbol{\beta}_0, \widehat{G}, \widehat{\mathrm{E}}), \widehat{\boldsymbol{\eta}} \right\rangle \\ &\leqslant \lambda(||\boldsymbol{\beta}_{0,-1}||_1 - ||\widehat{\boldsymbol{\beta}}_{-1}||_1) - \left\langle \boldsymbol{S}_n(\boldsymbol{\beta}_0, \widehat{G}, \widehat{\mathrm{E}}), \widehat{\boldsymbol{\eta}} \right\rangle \\ &\leqslant \lambda(||\boldsymbol{\beta}_{0,-1}||_1 - ||\widehat{\boldsymbol{\beta}}_{-1}||_1) + ||\boldsymbol{S}_n(\boldsymbol{\beta}_0, \widehat{G}, \widehat{\mathrm{E}})||_\infty ||\widehat{\boldsymbol{\eta}}||_1, \end{aligned} \tag{S22}$$

with probability at least $1 - \exp(-c_1 \log p)$, for some positive constant c_1 and all n sufficiently large. In the above, the second inequality uses (S21). Note that $||\widehat{\boldsymbol{\eta}}||_2 \leqslant ||\widehat{\boldsymbol{\eta}}||_1$. This implies that

$$c_0||\widehat{\boldsymbol{\eta}}||_2^2 \leqslant \left(c_1 h^2 + ||\boldsymbol{S}_n(\boldsymbol{\beta}_0, \widehat{G}, \widehat{\mathrm{E}})||_\infty \right) ||\widehat{\boldsymbol{\eta}}||_1 + \lambda \left(||\boldsymbol{\beta}_{0,-1}||_1 - ||\widehat{\boldsymbol{\beta}}_{-1}||_1 \right).$$

By Lemma A1, $\lambda/4 \geqslant ||\boldsymbol{S}_n(\boldsymbol{\beta}_0, \widehat{G}, \widehat{\mathrm{E}})||_\infty$ with probability at least $1 - \exp[-c_1 \log(p \vee n)]$, since $\sqrt{n^{-1}\log(p \vee n)} \leqslant c_0 h^{5/2} \leqslant c_0 h^2$ for some positive constant c_0. Let $\widehat{\boldsymbol{\eta}}_{\mathcal{S}}$ and $\widehat{\boldsymbol{\eta}}_{\mathcal{S}^C}$ be the sub-vectors of $\widehat{\boldsymbol{\eta}}$ on the support $\mathcal{S} = \{j : \beta_{0,j+1} \neq 0,\ j = 1, \cdots, p-1\}$, and $\mathcal{S}^C$, respectively. Then we have

$$c_0||\widehat{\boldsymbol{\eta}}||_2^2 \leqslant \frac{\lambda}{2}\left(||\widehat{\boldsymbol{\eta}}_{\mathcal{S}}||_1 + ||\widehat{\boldsymbol{\eta}}_{\mathcal{S}^C}||_1\right) + \lambda\left(||\widehat{\boldsymbol{\eta}}_{\mathcal{S}}||_1 - ||\widehat{\boldsymbol{\eta}}_{\mathcal{S}^C}||_1\right) \leqslant \frac{3\lambda}{2}||\widehat{\boldsymbol{\eta}}_{\mathcal{S}}||_1 - \frac{\lambda}{2}||\widehat{\boldsymbol{\eta}}_{\mathcal{S}^C}||_1, \tag{S23}$$

which implies that $||\widehat{\boldsymbol{\eta}}_{\mathcal{S}^C}||_1 \leqslant 3||\widehat{\boldsymbol{\eta}}_{\mathcal{S}}||_1$. Then (S23) implies that

$$c_0||\widehat{\boldsymbol{\eta}}||_2^2 \leqslant \left(c_1 h^2 + ||\boldsymbol{S}_n(\boldsymbol{\beta}_0, \widehat{G}, \widehat{\mathrm{E}})||_\infty + \lambda \right) ||\widehat{\boldsymbol{\eta}}||_1 \leqslant \frac{3\lambda}{2}||\widehat{\boldsymbol{\eta}}||_1 \leqslant 6\lambda||\widehat{\boldsymbol{\eta}}_{\mathcal{S}}||_1 \leqslant 6\lambda\sqrt{s}||\widehat{\boldsymbol{\eta}}||_2.$$

Hence $||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_2 = ||\widehat{\boldsymbol{\eta}}||_2 \leqslant \frac{6}{c_0}\lambda\sqrt{s}$. Since $||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_1 = ||\widehat{\boldsymbol{\eta}}||_1 \leqslant 4||\widehat{\boldsymbol{\eta}}_{\mathcal{S}}||_1 \leqslant 4\sqrt{s}||\widehat{\boldsymbol{\eta}}||_2$, the bound of $||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_1$ follows immediately. □

S5 Proofs of results in Section 3.2 of the main paper

Proof of Lemma 2. (1) To derive the uniform error bound for $\boldsymbol{d}_j(\widehat{\boldsymbol{\beta}}, \eta)$, $j = 2, \dots, p$, we first prove the following two results:

(i) $\boldsymbol{d}_{0j} = (\boldsymbol{\Omega}_{-(j-1),-(j-1)})^{-1}\boldsymbol{\Omega}_{-(j-1),(j-1)}$ is feasible in the sense that it satisfies the constraint of the Dantzig problem in (11) of the main paper with high probability, uniformly in $j = 2, \cdots, p$;

(ii) $\widehat{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^n [\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2 \widehat{\boldsymbol{x}}_{i,-1}\widehat{\boldsymbol{x}}_{i,-1}^T$ satisfies a restricted eigenvalue condition on the support of $\boldsymbol{\phi}_{0j} = \tau_{0j}^2\boldsymbol{\theta}_j$, denoted by $\mathcal{S}_{\boldsymbol{\phi}_j}$, with high probability, uniformly in $j = 2, \cdots, p$, as shown below in (S24).

To prove (i), the assumptions of Theorem 1 imply that $\sigma_x^2\sqrt{\frac{\log p}{n}} = O(h^{5/2}) = o(\eta)$. Lemma B8 implies that

$$P\left(\max_{2\leqslant j\leqslant p}\left|\left|\frac{1}{n}\sum_{i=1}^n [G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]^2 \widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\phi}_{0j}\widetilde{\boldsymbol{x}}_{i,-j*}\right|\right|_\infty \geqslant \eta/2\right) \leqslant \exp(-c_2\log p),$$

for some positive constant c_2, and all n sufficiently large.

Lemma B9 implies that

$$P\left(\max_{2\leqslant j\leqslant p}\left|\left|\frac{1}{n}\sum_{i=1}^n \left\{[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2 - [G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]^2\right\}\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\phi}_{0j}\widetilde{\boldsymbol{x}}_{i,-j*}\right|\right|_\infty \geqslant c_0h\right) \leqslant \exp(-c_1\log p),$$

for some positive constants c_0 and c_1, and all n sufficiently large. Lemma B10 implies there exist some positive constants c_0, c_1, such that for all n sufficiently large,

$$P\left(\max_{2\leqslant j\leqslant p}\left|\left|\frac{1}{n}\sum_{i=1}^n \left[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\right]^2(\widehat{\boldsymbol{x}}_{i,-1}\widehat{\boldsymbol{x}}_{i,-1}^T - \widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T)\boldsymbol{\phi}_{0j}\right|\right|_\infty \geqslant c_0\sqrt{s}h^2\right) \leqslant \exp(-c_1\log p).$$

Hence there exist some universal positive constants d_2 and c_1, such that for $\eta = d_2h$, for all

n sufficiently large,

$$
\begin{aligned}
&P\left(\max_{2\leqslant j\leqslant p}\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2\widehat{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\phi}_{0j}\widehat{\boldsymbol{x}}_{i,-j*}\Big|\Big|_\infty\geqslant\eta\right)\\
\leqslant&P\left(\max_{2\leqslant j\leqslant p}\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]^2\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\phi}_{0j}\widetilde{\boldsymbol{x}}_{i,-j*}\Big|\Big|_\infty\geqslant\eta/2\right)\\
&+P\left(\max_{2\leqslant j\leqslant p}\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\big\{[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2-[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]^2\big\}\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\phi}_{0j}\widetilde{\boldsymbol{x}}_{i,-j*}\Big|\Big|_\infty\geqslant\eta/4\right)\\
&+P\left(\max_{2\leqslant j\leqslant p}\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2(\widehat{\boldsymbol{x}}_{i,-1}\widehat{\boldsymbol{x}}_{i,-1}^T-\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T)\boldsymbol{\phi}_{0j}\Big|\Big|_\infty\geqslant\eta/4\right)\\
\leqslant&\exp(-c_1\log p),
\end{aligned}
$$

as $\sqrt{s}h^2\leqslant d_0h^2*\sqrt{nh^5}\leqslant d_0h\sqrt{nh^7}=o(\eta)$ for some positive constant d_0 by the assumptions of Theorem 1. Since $\boldsymbol{\phi}_{0j}=\Big(-(\boldsymbol{d}_{0j})_{1:(j-2)}^T,1,-(\boldsymbol{d}_{0j})_{(j-1):(p-2)}^T\Big)^T$ by Lemma A13, it implies that $\boldsymbol{d}_{0j}$ satisfies the constraint in (11) in Section 2.3 of the main paper, that is,

$$
\Big|\Big|n^{-1}\sum_{i=1}^{n}[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2(\widehat{x}_{i,j}-\widehat{\boldsymbol{x}}_{i,-j}^T\boldsymbol{d}_{0j})\widehat{\boldsymbol{x}}_{i,-j*}\Big|\Big|_\infty\leqslant\eta,
$$

with probability at least $1-\exp(-c_1\log p)$ uniformly in j, with $\eta=d_2h$, for some positive constants d_2, c_1 and all n sufficiently large. This ensures that $\boldsymbol{d}_{0j}$ is feasible for (11) with probability at least $1-\exp(-c_1\log p)$ uniformly in j, for all n sufficiently large. By the definition of $\boldsymbol{d}_j(\widehat{\boldsymbol{\beta}},\eta)$, we have

$$
P\Big(||\boldsymbol{d}_j(\widehat{\boldsymbol{\beta}},\eta)||_1\leqslant||\boldsymbol{d}_{0j}||_1\text{ uniformly in }j\Big)\geqslant 1-\exp(-c_1\log p),
$$

with $\eta=d_2h$, for some positive constants d_2, c_1 and all n sufficiently large.

Given (i), the event $\mathcal{E}_1=\Big\{||\boldsymbol{d}_j(\widehat{\boldsymbol{\beta}},\eta)||_1\leqslant||\boldsymbol{d}_{0j}||_1\text{ uniformly in }j\Big\}$ holds with probability at least $1-\exp(-c_1\log p)$ for some positive constant c_1 and all n sufficiently large. Define $\boldsymbol{w}_j=\boldsymbol{\phi}_j(\widehat{\boldsymbol{\beta}},\eta)-\boldsymbol{\phi}_{0j}$. Note that for any j such that $\boldsymbol{\phi}_{0j}=\boldsymbol{e}_{j-1}$, we have $\boldsymbol{d}_{0j}=\boldsymbol{0}_{p-2}$, by

Lemma A13, where $\boldsymbol{e}_{j-1}$ denotes the $(p-1)-$dimensional vector with the $(j-1)^{th}$ entry being one and all the other entries equal to zero, and $\mathbf{0}_{p-2}$ denotes the $(p-2)-$dimensional vector with all the entries equal to zero. On the event $\mathcal{E}_1$, if j is such that $\boldsymbol{\phi}_{0j} = \boldsymbol{e}_{j-1}$, then $||\boldsymbol{w}_j||_1 = 0$ (as we will have $||\boldsymbol{d}_j(\widehat{\boldsymbol{\beta}}, \eta)||_1 = ||\boldsymbol{d}_{0j}||_1 = 0$ for this case) and the results in Lemma 2-(1) always hold. Therefore, without loss of generality, we assume that $\boldsymbol{\phi}_{0j} \neq \boldsymbol{e}_{j-1}$ for any $j = 2, \ldots, p$. Recall that $\mathcal{S}_{\boldsymbol{\phi}_j}$ is the support set of $\boldsymbol{\phi}_{0j} = \tau_{0j}^2 \boldsymbol{\theta}_j$. On the event $\mathcal{E}_1$,

$$||\boldsymbol{w}_{j,\mathcal{S}_{\boldsymbol{\phi}_j}^C}||_1 = \left|\left|\left[\boldsymbol{\phi}_j(\widehat{\boldsymbol{\beta}}, \eta)\right]_{\mathcal{S}_{\boldsymbol{\phi}_j}^C}\right|\right|_1 \leqslant ||\boldsymbol{\phi}_{0j,\mathcal{S}_{\boldsymbol{\phi}_j}}||_1 - \left|\left|\left[\boldsymbol{\phi}_j(\widehat{\boldsymbol{\beta}}, \eta)\right]_{\mathcal{S}_{\boldsymbol{\phi}_j}}\right|\right|_1 \leqslant ||\boldsymbol{w}_{j,\mathcal{S}_{\boldsymbol{\phi}_j}}||_1,$$

where the first equality applies Lemma A13; the second last inequality applies $||\boldsymbol{\phi}_j(\widehat{\boldsymbol{\beta}}, \eta)||_1 \leqslant ||\boldsymbol{\phi}_{0j}||_1$; the last inequality applies $\left|\left|\boldsymbol{\phi}_{0j,\mathcal{S}_{\boldsymbol{\phi}_j}}\right|\right|_1 = \left|\left|[\boldsymbol{\phi}_j(\widehat{\boldsymbol{\beta}}, \eta)]_{\mathcal{S}_{\boldsymbol{\phi}_j}} - \boldsymbol{w}_{j,\mathcal{S}_{\boldsymbol{\phi}_j}}\right|\right|_1 \leqslant \left|\left|[\boldsymbol{\phi}_j(\widehat{\boldsymbol{\beta}}, \eta)]_{\mathcal{S}_{\boldsymbol{\phi}_j}}\right|\right| + \left|\left|\boldsymbol{w}_{j,\mathcal{S}_{\boldsymbol{\phi}_j}}\right|\right|_1$. Denote the set $\mathcal{V}_{2j} = \{\boldsymbol{v} = (v_1, \cdots, v_{p-1})^T : ||\boldsymbol{v}_{\mathcal{S}_{\boldsymbol{\phi}_j}^C}||_1 \leqslant ||\boldsymbol{v}_{\mathcal{S}_{\boldsymbol{\phi}_j}}||_1, ||\boldsymbol{v}||_2 = 1, v_{j-1} = 0\}$, for any $j = 2, \ldots, p$. We observe that on the event $\mathcal{E}_1$, $\frac{\boldsymbol{w}_j}{||\boldsymbol{w}_j||_2} \in \mathcal{V}_{2j}$, for any $j = 2, \ldots, p$. In the next step, we will prove

$$P\left(\min_{2 \leqslant j \leqslant p} \inf_{\boldsymbol{v} \in \mathcal{V}_{2j}} \boldsymbol{v}^T \widehat{\boldsymbol{\Omega}} \boldsymbol{v} \leqslant \frac{\xi_2}{2}\right) \leqslant \exp(-d_0 \log p), \tag{S24}$$

for some positive constant d_0 and all n sufficiently large.

To prove (S24), Assumption (A2)-(a) indicates that

$$\begin{aligned}\min_{2 \leqslant j \leqslant p} \inf_{\boldsymbol{v} \in \mathcal{V}_{2j}} \boldsymbol{v}^T \widehat{\boldsymbol{\Omega}} \boldsymbol{v} &\geqslant \min_{2 \leqslant j \leqslant p} \inf_{\boldsymbol{v} \in \mathcal{V}_{2j}} \boldsymbol{v}^T \boldsymbol{\Omega} \boldsymbol{v} - \max_{2 \leqslant j \leqslant p} \sup_{\boldsymbol{v} \in \mathcal{V}_{2j}} \boldsymbol{v}^T \{\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\} \boldsymbol{v} \\ &\geqslant \xi_2 - \max_{2 \leqslant j \leqslant p} \sup_{\boldsymbol{v} \in \mathcal{V}_{2j}} \boldsymbol{v}^T \{\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\} \boldsymbol{v},\end{aligned}$$

where ξ_2 is the positive constant defined in Assumption (A2)-(a). Note that

$$\max_{2\leqslant j\leqslant p}\sup_{\boldsymbol{v}\in\mathcal{V}_{2j}}\left|\boldsymbol{v}^T\{\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}\}\boldsymbol{v}\right|\leqslant\max_{2\leqslant j\leqslant p}\sup_{\boldsymbol{v}\in\mathcal{V}_{2j}}\left|\boldsymbol{v}^T\Big(\frac{1}{n}\sum_{i=1}^n[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2[\widehat{\boldsymbol{x}}_i\widehat{\boldsymbol{x}}_i^T-\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T]\Big)\boldsymbol{v}\right|$$

$$+\max_{2\leqslant j\leqslant p}\sup_{\boldsymbol{v}\in\mathcal{V}_{2j}}\left|\frac{1}{n}\sum_{i=1}^n\Big\{[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2-[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]^2\Big\}(\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{v})^2\right|$$

$$+\max_{2\leqslant j\leqslant p}\sup_{\boldsymbol{v}\in\mathcal{V}_{2j}}\left|\boldsymbol{v}^T\Big(\frac{1}{n}\sum_{i=1}^n[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]^2\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T-\boldsymbol{\Omega}\Big)\boldsymbol{v}\right|.$$

Note that for any $\boldsymbol{v}\in\mathcal{V}_{2j}$, we have that $||\boldsymbol{v}||_1\leqslant 2||\boldsymbol{v}_{\mathcal{S}_{\phi_j}}||_1\leqslant 2\sqrt{s_j}||\boldsymbol{v}_{\mathcal{S}_{\phi_j}}||_2\leqslant 2\sqrt{s_j}\leqslant 2\sqrt{\widetilde{s}}$, where $s_j=||\boldsymbol{d}_{0j}||_0$, and $\widetilde{s}=\max_{2\leqslant j\leqslant p}s_j$. Hence the proof of Lemma B10 implies that $\max_{2\leqslant j\leqslant p}\sup_{\boldsymbol{v}\in\mathcal{V}_{2j}}\left|\boldsymbol{v}^T\left(\frac{1}{n}\sum_{i=1}^n[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2[\widehat{\boldsymbol{x}}_{i,-1}\widehat{\boldsymbol{x}}_{i,-1}^T-\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T]\right)\boldsymbol{v}\right|\leqslant c\widetilde{s}sh^4\leqslant c_0\widetilde{s}h^4*nh^5\leqslant c_0h^2$ with probability at least $1-\exp(-c_1\log p)$, by the assumptions of Lemma 2, for some positive constants c, c_0, c_1, and all n sufficiently large. The proof of Lemma B9 implies that with probability at least $1-\exp(-c_1\log p)$,

$$\max_{2\leqslant j\leqslant p}\sup_{\boldsymbol{v}\in\mathcal{V}_{2j}}\left|\frac{1}{n}\sum_{i=1}^n\Big\{[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2-[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]^2\Big\}(\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{v})^2\right|\leqslant c_0h(1+\sqrt{n^{-1}\widetilde{s}\log p}),$$

for some positive constants c_0, c_1, and all n sufficiently large. Note that $\sqrt{n^{-1}\widetilde{s}\log p}\leqslant 1$ by the assumptions of Lemma 2. Similarly as Lemma A2, $(2A_i-1)G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\widetilde{\boldsymbol{x}}_{i,-1}$ is sub-Gaussian with variance proxy no larger than $b^2\sigma_x^2$, where b is the positive constant defined in Assumption (A1)-(b). Similarly as Lemma B2, we have

$$P\left(\max_{2\leqslant j\leqslant p}\sup_{\boldsymbol{v}\in\mathcal{V}_{2j}}\left|\boldsymbol{v}^T\Big(\frac{1}{n}\sum_{i=1}^n[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]^2\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T-\boldsymbol{\Omega}\Big)\boldsymbol{v}\right|\geqslant c_0\sigma_x^2\sqrt{\frac{\widetilde{s}\log p}{n}}\right)\leqslant\exp(-c_1\widetilde{s}\log p),$$

where $\widetilde{s}=\max_{2\leqslant j\leqslant p}||\boldsymbol{d}_{0j}||_0=\max_{2\leqslant j\leqslant p}||\boldsymbol{\phi}_{0j}||_0-1$, by Lemma A13, for some positive constants c_0, c_1, and all n sufficiently large. Hence there exist some positive constants c_0, c_1, such that for all n sufficiently large,

$$P\left(\max_{2\leqslant j\leqslant p}\sup_{\boldsymbol{v}\in\mathcal{V}_{2j}}\left|\boldsymbol{v}^T\{\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}\}\boldsymbol{v}\right|\geqslant c_0\big(h+\sqrt{n^{-1}\widetilde{s}\log p}\big)\right)\leqslant\exp(-c_1\log p).$$

Note that $\sqrt{\frac{\widetilde{s}\log p}{n}}\leqslant c_2\sqrt{\widetilde{s}h^5}\leqslant c_2h\sqrt{\widetilde{s}h^3}=o(h)$, for some positive constant c_2, by the assumptions of Theorem 1 and Lemma 2. Hence we conclude

$$\begin{aligned}P\left(\min_{2\leqslant j\leqslant p}\inf_{\boldsymbol{v}\in\mathcal{V}_{2j}}\boldsymbol{v}^T\widehat{\boldsymbol{\Omega}}\boldsymbol{v}\leqslant\frac{\xi_2}{2}\right)\leqslant&P\Big(\xi_2-\max_{2\leqslant j\leqslant p}\sup_{\boldsymbol{v}\in\mathcal{V}_{2j}}\boldsymbol{v}^T\{\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}\}\boldsymbol{v}\leqslant\frac{\xi_2}{2}\Big)\\ \leqslant&P\left(\max_{2\leqslant j\leqslant p}\sup_{\boldsymbol{v}\in\mathcal{V}_{2j}}\left|\boldsymbol{v}^T\{\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}\}\boldsymbol{v}\right|\geqslant\frac{\xi_2}{2}\right)\leqslant\exp(-c_1\log p),\end{aligned}$$

for some positive constant c_1 and all n sufficiently large. It proves (S24), i.e., (ii).

Given (i) and (ii), the event

$$\mathcal{E}=\left\{\min_{2\leqslant j\leqslant p}\inf_{\boldsymbol{v}\in\mathcal{V}_{2j}}\boldsymbol{v}^T\widehat{\boldsymbol{\Omega}}\boldsymbol{v}\geqslant\frac{\xi_2}{2},\text{ and }||\boldsymbol{d}_j(\widehat{\boldsymbol{\beta}},\eta)||_1\leqslant||\boldsymbol{d}_{0j}||_1\text{ uniformly in }j\right\},$$

holds with probability at least $1-\exp(-c_1\log p)$ for some positive constant c_1 and all n sufficiently large. On the event $\mathcal{E}$, $\boldsymbol{d}_{0j}$ and $\boldsymbol{d}_j(\widehat{\boldsymbol{\beta}},\eta)$ both satisfy the constraint of the Dantzig problem in (11) of the main paper with high probability. Then we have

$$\begin{aligned}||\widehat{\boldsymbol{\Omega}}\boldsymbol{w}_j||_\infty\leqslant&\Big|\Big|\frac{1}{n}\sum_{i=1}^n[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2\big[\widehat{x}_{i,j}-\widehat{\boldsymbol{x}}_{i,-j*}^T\boldsymbol{d}_j(\widehat{\boldsymbol{\beta}},\eta)\big]\widehat{\boldsymbol{x}}_{i,-j*}\Big|\Big|_\infty\\&+\Big|\Big|\frac{1}{n}\sum_{i=1}^n[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2\big[\widehat{x}_{i,j}-\widehat{\boldsymbol{x}}_{i,-j*}^T\boldsymbol{d}_{0j}\big]\widehat{\boldsymbol{x}}_{i,-j*}\Big|\Big|_\infty\leqslant2\eta.\end{aligned}$$

Note that on the event $\mathcal{E}$, $\frac{\boldsymbol{w}_j}{||\boldsymbol{w}_j||_2}\in\mathcal{V}_{2j}$ holds uniformly in j. On the event $\mathcal{E}$, we have

$$||\boldsymbol{w}_j||_1^2\leqslant4s_j||\boldsymbol{w}_j||_2^2\leqslant\frac{8s_j\boldsymbol{w}_j^T\widehat{\boldsymbol{\Omega}}\boldsymbol{w}_j}{\xi_2}\leqslant\frac{8s_j||\boldsymbol{w}_j||_1||\widehat{\boldsymbol{\Omega}}\boldsymbol{w}_j||_\infty}{\xi_2}\leqslant\frac{16s_j\eta||\boldsymbol{w}_j||_1}{\xi_2}\leqslant\frac{32s_j^{3/2}\eta||\boldsymbol{w}_j||_2}{\xi_2},$$

uniformly in j, where the second last inequality applies the above result. It hence implies that $P\Big(||\boldsymbol{w}_j||_1 \leqslant \frac{16 s_j \eta}{\xi_2} \text{ uniformly in } j\Big) \geqslant P(\mathcal{E}) \geqslant 1 - \exp(-c_1 \log p)$, and $P\Big(||\boldsymbol{w}_j||_2 \leqslant \frac{8\sqrt{s_j}\eta}{\xi_2} \text{ uniformly in } j\Big) \geqslant 1 - \exp(-c_1 \log p)$, for some positive constant c_1 and all n sufficiently large. Note that $||\boldsymbol{w}_j||_1 = ||\boldsymbol{d}_j(\widehat{\boldsymbol{\beta}}, \eta) - \boldsymbol{d}_{0j}||_1$ and $||\boldsymbol{w}_j||_2 = ||\boldsymbol{d}_j(\widehat{\boldsymbol{\beta}}, \eta) - \boldsymbol{d}_{0j}||_2$, (1) is proved.

(2) Recall that $\tau_{0j}^2 = \mathrm{E}\big\{[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]^2 \widetilde{x}_{i,j}\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\phi}_{0j}\big\}$. We have

$$
\begin{aligned}
\Big|\tau_{0j}^2 - \tau_j^2(\widehat{\boldsymbol{\beta}}, \eta)\Big| \leqslant & \left|\frac{1}{n}\sum_{i=1}^n \Big\{[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]^2 - [\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2\Big\}\, \widetilde{x}_{i,j}\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\phi}_{0j}\right| \\
& + \left|\frac{1}{n}\sum_{i=1}^n [\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2[\widehat{x}_{i,j}\widehat{\boldsymbol{x}}_{i,-1}^T - \widetilde{x}_{i,j}\widetilde{\boldsymbol{x}}_{i,-1}^T]\boldsymbol{\phi}_{0j}\right| \\
& + \left|\frac{1}{n}\sum_{i=1}^n [\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2\widehat{x}_{i,j}\widehat{\boldsymbol{x}}_{i,-1}^T\boldsymbol{w}_j\right| \\
& + \left|\frac{1}{n}\sum_{i=1}^n [G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]^2\widetilde{x}_{i,j}\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\phi}_{0j} - \tau_{0j}^2\right| \\
\triangleq & |I_{j1}| + |I_{j2}| + |I_{j3}| + |I_{j4}|,
\end{aligned}
$$

where the definition of I_{jk} is clear from the context, and $\boldsymbol{w}_j = \boldsymbol{\phi}_j(\widehat{\boldsymbol{\beta}}, \eta) - \boldsymbol{\phi}_{0j}$. Lemma B9 implies that there exist some positive constants c_0, c_1, such that for all n sufficiently large,

$$
\begin{aligned}
& P\left(|I_{j1}| \geqslant c_0\sqrt{s_j}\eta \text{ uniformly in } j\right) \\
\leqslant & P\left(\left|\left|\frac{1}{n}\sum_{i=1}^n \big\{[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2 - [G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]^2\big\}\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T\right|\right|_\infty \geqslant c_0\eta\right) \\
\leqslant & \exp(-c_1\log p).
\end{aligned}
$$

Lemma B10 implies that $P(|I_{j2}| \geqslant c_0\sqrt{s_j}\eta \text{ uniformly in } j) \leqslant \exp(-c_1\log p)$, for some positive constants c_0, c_1 and all n sufficiently large, since $\sqrt{s}h^2 \leqslant d_0h^2\sqrt{nh^5} = d_0h\sqrt{nh^7} = o(\eta)$, for some positive constant d_0 by the assumptions of Theorem 1 and Lemma 2. Lemma A6 and Assumption (A5)-(a) imply $\max_{1\leqslant i\leqslant n}[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2 \leqslant c_0(b+h)^2 \leqslant 2c_0b^2$ with probability at least $1 - \exp[-c_1\log(p\vee n)]$, for positive constants c_0, c_1, b, and all n sufficiently large. Then there are some universal positive constants c_0 and c_1 such that for all n sufficiently large,

$$
\begin{aligned}
|I_{j3}| \leqslant & \max_{1\leqslant i\leqslant n}[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2 * \frac{1}{n}\sum_{i=1}^n |\widetilde{x}_{i,j}\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{w}_j| + \frac{1}{n}\sum_{i=1}^n [\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2|(\widehat{x}_{i,j}\widehat{\boldsymbol{x}}_{i,-1}^T - \widetilde{x}_{i,j}\widetilde{\boldsymbol{x}}_{i,-1}^T)\boldsymbol{w}_j| \\
\leqslant & c_0\sqrt{\frac{1}{n}\sum_{i=1}^n \widetilde{x}_{i,j}^2}\sqrt{\frac{1}{n}\sum_{i=1}^n(\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{w}_j)^2} + c_0||\boldsymbol{w}_j||_1 * \frac{1}{n}\sum_{i=1}^n[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2||\widehat{\boldsymbol{x}}_{i,-1}\widehat{\boldsymbol{x}}_{i,-1}^T - \widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T||_\infty \\
\leqslant & c_0\sqrt{\frac{1}{n}\sum_{i=1}^n \widetilde{x}_{i,j}^2}\sqrt{\boldsymbol{w}_j^T\Big(\frac{1}{n}\sum_{i=1}^n \widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T\Big)\boldsymbol{w}_j} \\
& + c_0\eta\sqrt{\widetilde{s}} * \frac{1}{n}\sum_{i=1}^n[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2||\widehat{\boldsymbol{x}}_{i,-1}\widehat{\boldsymbol{x}}_{i,-1}^T - \widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T||_\infty,
\end{aligned}
$$

uniformly in j with probability at least $1-\exp[-c_1\log(p\vee n)]$, applying the bound of $||\boldsymbol{w}_j||_1$ derived in the proof of (1). Lemma B1 and Lemma B2 imply that for some universal positive constants c_0, c_1, and all n sufficiently large,

$$
\begin{aligned}
P\left(\max_{2\leqslant j\leqslant p}\frac{1}{n}\sum_{i=1}^n\widetilde{x}_{i,j}^2 \geqslant \max_{2\leqslant j\leqslant p}\mathrm{E}(\widetilde{x}_{i,j}^2) + c_0\sigma_x^2\sqrt{\frac{\log p}{n}}\right) &\leqslant \exp(-c_1\log p), \\
P\left(\Big|\Big|\frac{1}{n}\sum_{i=1}^n\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T - \mathrm{E}(\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T)\Big|\Big|_\infty \geqslant c_0\sigma_x^2\sqrt{\frac{\log p}{n}}\right) &\leqslant \exp(-c_1\log p), \\
P\Bigg(\Big|\boldsymbol{w}_j^T\Big(\frac{1}{n}\sum_{i=1}^n\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T - \mathrm{E}(\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T)\Big)\boldsymbol{w}_j\Big| & \\
\geqslant c_0\sigma_x^2||\boldsymbol{w}_j||_2^2\sqrt{\frac{\log p}{n}},\ \text{uniformly in } j\Bigg) &\leqslant \exp(-c_1\log p),
\end{aligned}
$$

with $\boldsymbol{w}_j = \boldsymbol{\phi}_j(\widehat{\boldsymbol{\beta}},\eta) - \boldsymbol{\phi}_{0j}$. Since $\widetilde{x}_{i,j}$ is sub-Gaussian with variance proxy at most $2\sigma_x^2$ uniformly in j by Lemma A2, we have $\max_{2\leqslant j\leqslant p}\mathrm{E}(\widetilde{x}_{i,j}^2) \leqslant 2\sigma_x^2$. Note that $\mathrm{E}\left[\mathrm{Cov}(\boldsymbol{x}_{-1}|\boldsymbol{x}^T\boldsymbol{\beta}_0)\right] = \mathrm{E}(\widetilde{\boldsymbol{x}}_{-1}\widetilde{\boldsymbol{x}}_{-1}^T)$. Assumption (A2)-(a) thus implies that ξ_1 is the largest eigenvalue of $\mathrm{E}(\widetilde{\boldsymbol{x}}_{-1}\widetilde{\boldsymbol{x}}_{-1}^T)$. The results in part (1) of the lemma indicate that there exist some positive constants c_0, c_1, such that for all n sufficiently large,

$$
\boldsymbol{w}_j^T\Big(\frac{1}{n}\sum_{i=1}^n\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T\Big)\boldsymbol{w}_j \leqslant 2\boldsymbol{w}_j^T\mathrm{E}(\widetilde{\boldsymbol{x}}_{-1}\widetilde{\boldsymbol{x}}_{-1}^T)\boldsymbol{w}_j \leqslant 2\xi_1||\boldsymbol{w}_j||_2^2 \leqslant c_0\xi_1 s_j\eta^2 \text{ uniformly in } j,
$$

with probability at least $1 - \exp(-c_1 \log p)$, where the first inequality applies Lemma B1. Lemma B10 implies that there exist some positive constants c_0, c_1, such that for all n sufficiently large,

$$\frac{1}{n}\sum_{i=1}^{n}[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2||\widehat{\boldsymbol{x}}_{i,-1}\widehat{\boldsymbol{x}}_{i,-1}^T - \widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T||_\infty \leqslant c_0\sqrt{s}h^2,$$

with probability at least $1 - \exp(-c_1 \log p)$. The assumptions of Theorem 1 and Lemma 2 imply that $\sqrt{s\widetilde{s}}h^2\eta = \eta\sqrt{sh^3}\sqrt{\widetilde{s}h} \leqslant d_0\eta\sqrt{nh^8} = o(\eta)$, for some positive constant d_0, since $nh^6 = O(1)$. It follows that $|I_{j3}| \leqslant c_0\sqrt{s_j}\eta$ uniformly in j with probability at least $1 - \exp(-c_1 \log p)$ for universal positive constants c_0, c_1, and all n sufficiently large,

To uniformly bound $|I_{4j}|$, as in the proof of part (1) of the lemma, we observe that $G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\widetilde{x}_{i,j}$ and $G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0) * \widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\phi}_{0j}$ are both sub-Gaussian with variance proxy at most $c_1 b^2\sigma_x^2$. There exist some positive constants c_0, c, such that for all n sufficiently large,

$$\begin{aligned}P\left(|I_{j4}| \geqslant c_0\sqrt{s_j}\eta \text{ uniformly in } j\right) \leqslant & \sum_{j=2}^{p} P\left(\left|\frac{1}{n}\sum_{i=1}^{n}[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]^2\widetilde{x}_{i,j}\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\phi}_{0j} - \tau_{0j}^2\right| \geqslant c_0\sqrt{s_j}\eta\right)\\ \leqslant & p\exp(-cn\eta^2).\end{aligned}$$

Since $\log p \leqslant d_0 nh^5 = o(n\eta^2)$ for some positive constant d_0, there exist some positive constants c_0, c_1, such that for all n sufficiently large,

$$P\Big(|\tau_{0j}^2 - \tau_j^2(\widehat{\boldsymbol{\beta}}, \eta)| \leqslant c_0\sqrt{s_j}\eta \text{ uniformly in } j\Big) \leqslant \exp(-c_1 \log p).$$

The first result of part (2) of Lemma 2 is proved. Note that

$$
\begin{aligned}
&P\Big(|\tau_{0j}^{-2} - \tau_j^{-2}(\widehat{\boldsymbol{\beta}}, \eta)| \geqslant c_0\sqrt{s_j}\eta \text{ uniformly in } j\Big)\\
=&P\Big(\tau_{0j}^{-2} * \tau_j^{-2}(\widehat{\boldsymbol{\beta}}, \eta)|\tau_{0j}^2 - \tau_j^2(\widehat{\boldsymbol{\beta}}, \eta)| \geqslant c_0\sqrt{s_j}\eta \text{ uniformly in } j\Big)\\
\leqslant&P\Big(|\tau_{0j}^2 - \tau_j^2(\widehat{\boldsymbol{\beta}}, \eta)| \leqslant c_0\xi_2^2\sqrt{s_j}\eta/2 \text{ uniformly in } j\Big) + P\Big(\max_{2\leqslant j\leqslant p} \tau_{0j}^{-2} * \tau_j^{-2}(\widehat{\boldsymbol{\beta}}, \eta) \leqslant 2\xi_2^{-2}\Big)\\
\leqslant&\exp(-c_1\log p) + P\Big(\max_{2\leqslant j\leqslant p} \tau_{0j}^{-2} * \tau_j^{-2}(\widehat{\boldsymbol{\beta}}, \eta) \leqslant 2\xi_2^{-2}\Big),
\end{aligned}
$$

for some positive constants c_0 c_1, and all n sufficiently large. Lemma A13 implies that $\xi_2 \leqslant \tau_{0j}^2 \leqslant b^2\xi_1$, uniformly in j. By the first result of part (2) of the lemma, we know that $\xi_2/2 \leqslant \tau_j^2(\widehat{\boldsymbol{\beta}}, \eta) \leqslant b^2\xi_1/2$ with probability at least $1 - \exp(-c_1\log p)$, for some positive constant c_1 and all n sufficiently large. Then we have $P\Big(\max_{2\leqslant j\leqslant p} \tau_{0j}^{-2} * \tau_j^{-2}(\widehat{\boldsymbol{\beta}}, \eta) \leqslant 2\xi_2^{-2}\Big) \leqslant \exp(-c_1\log p)$ for some positive constant c_1 and all n sufficiently large. Hence the second result of part (2) of Lemma 2 is proved.

(3) Observe that

$$
||\boldsymbol{\theta}_j(\widehat{\boldsymbol{\beta}}, \eta) - \boldsymbol{\theta}_j||_1 = ||\tau_j^{-2}(\widehat{\boldsymbol{\beta}}, \eta)\boldsymbol{\phi}_j(\widehat{\boldsymbol{\beta}}, \eta) - \tau_{0j}^{-2}\boldsymbol{\phi}_{0j}||_1 \leqslant \tau_j^{-2}(\widehat{\boldsymbol{\beta}}, \eta)||\boldsymbol{w}_j||_1 + |\tau_j^{-2}(\widehat{\boldsymbol{\beta}}, \eta) - \tau_{0j}^{-2}| * ||\boldsymbol{\phi}_{0j}||_1,
$$

where $\boldsymbol{w}_j = \boldsymbol{\phi}_j(\widehat{\boldsymbol{\beta}}, \eta) - \boldsymbol{\phi}_{0j}$. Lemma A13 implies $||\boldsymbol{\phi}_{0j}||_1 = \tau_{0j}^2||\boldsymbol{\theta}_j||_1 \leqslant \sqrt{s_j}\tau_{0j}^2||\boldsymbol{\theta}_j||_2 \leqslant b\sqrt{s_j}\xi_1\xi_2^{-1}$ uniformly in j, and $\max_{2\leqslant j\leqslant p} \tau_j^{-2}(\widehat{\boldsymbol{\beta}}, \eta) \leqslant \max_{2\leqslant j\leqslant p} \tau_{0j}^{-2} + c_0\eta \leqslant 2\xi_2^{-2}$, with probability at least $1 - \exp(-c_1\log p)$, for some positive constants c_0, c_1 and all n sufficiently large. Results in (1) and (2) imply that there exist some positive constants d_0, d_1, d_2, such that for all n sufficiently large,

$$
||\boldsymbol{\theta}_j(\widehat{\boldsymbol{\beta}}, \eta) - \boldsymbol{\theta}_j||_1 \leqslant d_1 s_j\eta\xi_2^{-2} + d_1 s_j\eta * b^2\xi_1\xi_2^{-1} \leqslant d_0 s_j\eta \text{ uniformly in } j
$$

with probability at least $1 - \exp(-d_2\log p)$. Similar proofs can be applied for the uniform bound of $||\boldsymbol{\theta}_j(\widehat{\boldsymbol{\beta}}, \eta) - \boldsymbol{\theta}_j||_2$. $\square$

Proof of Theorem 2. Recall that

$$\boldsymbol{S}_n(\widehat{\boldsymbol{\beta}}, \widehat{G}, \widehat{\mathrm{E}}) = -n^{-1}\sum_{i=1}^{n}\Big[\widetilde{\epsilon}_i + G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) - \widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\Big]\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{x}}_{i,-1},$$

$$\boldsymbol{S}_n(\boldsymbol{\beta}_0, G, \mathrm{E}) = -n^{-1}\sum_{i=1}^{n}\widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\widetilde{\boldsymbol{x}}_{i,-1},$$

where $\widetilde{\boldsymbol{x}}_{i,-1} = \boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)$, $\widehat{\boldsymbol{x}}_{i,-1} = \boldsymbol{x}_{i,-1} - \widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})$, and $G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) = f_0'(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)$. Recall that the debiased estimator $\widetilde{\boldsymbol{\beta}}_{-1} = \widehat{\boldsymbol{\beta}}_{-1} - \widehat{\boldsymbol{\Theta}}\boldsymbol{S}_n(\widehat{\boldsymbol{\beta}}, \widehat{G}, \widehat{\mathrm{E}})$. We have

$$\begin{aligned}
&\sqrt{n}(\widetilde{\boldsymbol{\beta}}_{-1} - \boldsymbol{\beta}_{0,-1})\\
=&\sqrt{n}(\widehat{\boldsymbol{\beta}}_{-1} - \boldsymbol{\beta}_{0,-1}) - \sqrt{n}\widehat{\boldsymbol{\Theta}}^T\boldsymbol{S}_n(\widehat{\boldsymbol{\beta}}, \widehat{G}, \widehat{\mathrm{E}})\\
=&-\sqrt{n}\widehat{\boldsymbol{\Theta}}^T\boldsymbol{S}_n(\boldsymbol{\beta}_0, G, \mathrm{E}) + \sqrt{n}(\widehat{\boldsymbol{\beta}}_{-1} - \boldsymbol{\beta}_{0,-1}) - \sqrt{n}\widehat{\boldsymbol{\Theta}}^T\big[\boldsymbol{S}_n(\widehat{\boldsymbol{\beta}}, \widehat{G}, \widehat{\mathrm{E}}) - \boldsymbol{S}_n(\boldsymbol{\beta}_0, G, \mathrm{E})\big]\\
=&-\sqrt{n}\widehat{\boldsymbol{\Theta}}^T\boldsymbol{S}_n(\boldsymbol{\beta}_0, G, \mathrm{E}) + \sqrt{n}(\boldsymbol{I}_{p-1} - \widehat{\boldsymbol{\Theta}}^T\boldsymbol{J}_1)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\\
&-\sqrt{n}\widehat{\boldsymbol{\Theta}}^T\big[\boldsymbol{S}_n(\widehat{\boldsymbol{\beta}}, \widehat{G}, \widehat{\mathrm{E}}) - \boldsymbol{S}_n(\boldsymbol{\beta}_0, G, \mathrm{E}) - \boldsymbol{J}_1(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\big]\\
\triangleq&\boldsymbol{A}_{n1} + \boldsymbol{A}_{n2} + \boldsymbol{A}_{n3},
\end{aligned}$$

where the definition of $\boldsymbol{A}_{ni}$ is clear from the context, $\boldsymbol{I}_{p-1}$ is the $(p-1)-$dimensional identity matrix, $\boldsymbol{J}_1 = n^{-1}\sum_{i=1}^{n}\big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\big]^2\widehat{\boldsymbol{x}}_{i,-1}\widehat{\boldsymbol{x}}_{i,-1}^T$ is the leading term in the approximation to $\nabla\boldsymbol{S}_n(\boldsymbol{\beta}_0, G, \mathrm{E})$. Let $\Delta_{n,p} = sh^3\sqrt{n} + \widetilde{s}h\sqrt{\log p}$. To prove the theorem, we will verify:

(1) The j^{th} component of $\boldsymbol{A}_{n1}$, denoted as $\boldsymbol{A}_{n1j}$, is approximately normal, for any $2 \leqslant j \leqslant p$;

(2) $P(||\boldsymbol{A}_{n2}||_\infty \geqslant c_0\Delta_{n,p}) \leqslant \exp(-c_1\log p)$;

(3) $P(||\boldsymbol{A}_{n3}||_\infty \geqslant c_0\Delta_{n,p}) \leqslant \exp(-c_1\log p)$;

for some positive constants c_0 and c_1, and all n sufficiently large.

To prove (1), we observe that $-\sqrt{n}\boldsymbol{e}_{j-1}^T\boldsymbol{\Theta}^T\boldsymbol{S}_n(\boldsymbol{\beta}_0, G, \mathrm{E}) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{e}_{j-1}^T\boldsymbol{\Theta}^T\boldsymbol{\Lambda}\boldsymbol{\Theta}\boldsymbol{e}_{j-1})$ by the central limit theorem, with $\boldsymbol{\Lambda} = \mathrm{E}\Big\{\big[\widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\big]^2\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T\Big\}$, and $\boldsymbol{\Theta}\boldsymbol{e}_{j-1} = \boldsymbol{\theta}_j$, where

$\boldsymbol{e}_{j-1}$ is the $(j-1)^{th}$ column of the identity matrix $\boldsymbol{I}_{p-1}$. It suffices to prove that

$$P\left(\max_{2\leqslant j\leqslant p}\left|\sqrt{n}(\widehat{\boldsymbol{\theta}}_j-\boldsymbol{\theta}_j)^T\boldsymbol{S}_n(\boldsymbol{\beta}_0,G,\mathrm{E})\right|\geqslant c_0\Delta_{n,p}\right)\leqslant \exp(-c_1\log p),$$

for some positive constants c_0, c_1, and all n sufficiently large.

Note that $2(2A_i-1)$ is a Rademacher sequence and independent of $(\epsilon_i,\boldsymbol{x}_i)$. We know that

$$\sqrt{n}\boldsymbol{S}_n(\boldsymbol{\beta}_0,G,\mathrm{E})=n^{-1/2}\sum_{i=1}^{n}2(2A_i-1)[\epsilon_i+g(\boldsymbol{x}_i)]G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\widetilde{\boldsymbol{x}}_{i,-1},$$

has mean zero; $2(2A_i-1)[\epsilon_i+g(\boldsymbol{x}_i)]G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)$ and $\widetilde{\boldsymbol{x}}_{i,-1}$ are both sub-Gaussian. Note that

$$\begin{aligned}\mathrm{E}\left[\boldsymbol{S}_n(\boldsymbol{\beta}_0,G,\mathrm{E})\right]=&\mathrm{E}_{\boldsymbol{x}_i^T\boldsymbol{\beta}_0}\left[\boldsymbol{S}_n(\boldsymbol{\beta}_0,G,\mathrm{E})|\boldsymbol{x}_i^T\boldsymbol{\beta}_0\right]\\=&\mathrm{E}_{\boldsymbol{x}_i^T\boldsymbol{\beta}_0}\left\{2G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\mathrm{E}\Big\{(2A_i-1)[\epsilon_i+g(\boldsymbol{x}_i)]\widetilde{\boldsymbol{x}}_{i,-1}\Big|\boldsymbol{x}_i^T\boldsymbol{\beta}_0\Big\}\right\}\\=&\mathbf{0}_{p-1},\end{aligned}$$

where $\mathbf{0}_{p-1}$ is a $(p-1)-$dimensional vector with all entries 0. Hence Lemma B1 implies that

$$P\Big(\big|\big|\sqrt{n}\boldsymbol{S}_n(\boldsymbol{\beta}_0,G,\mathrm{E})\big|\big|_\infty\geqslant c_0\sqrt{\log p}\Big)\leqslant\exp(-c_1\log p),$$

for some positive constants c_0, c_1 and all n sufficiently large. Hence, according to Lemma 2, there exist some positive constants c_0, c_1, c_2, such that for all n sufficiently large,

$$\begin{aligned}&P\left(\max_{2\leqslant j\leqslant p}\left|\sqrt{n}(\widehat{\boldsymbol{\theta}}_j-\boldsymbol{\theta}_j)^T\boldsymbol{S}_n(\boldsymbol{\beta}_0,G,\mathrm{E})\right|\geqslant c_0c_2\widetilde{s}h\sqrt{\log p}\right)\\\leqslant&P\Big(\big|\big|\sqrt{n}\boldsymbol{S}_n(\boldsymbol{\beta}_0,G,\mathrm{E})\big|\big|_\infty\geqslant c_0\sqrt{\log p}\Big)+\sum_{j=2}^{p}P\Big(||\widehat{\boldsymbol{\theta}}_j-\boldsymbol{\theta}_j||_1\geqslant c_2\widetilde{s}h\Big)\\\leqslant&\exp(-c_1\log p),\end{aligned}$$

which completes the proof for (1).

To prove (2), recall the definition of $\widehat{\boldsymbol{\theta}}_j$ in (14) of the main paper. Lemma A13 and Lemma 2 imply that there exist some positive constants c_0, c_1, c_2, such that for all n sufficiently large,

$$\begin{aligned}
&\Big\|n^{-1}\sum_{i=1}^{n}[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2\widehat{\boldsymbol{x}}_i^T\widehat{\boldsymbol{\theta}}_j\widehat{\boldsymbol{x}}_{i,-j*}\Big\|_\infty\\
\leqslant&\Big\|n^{-1}\sum_{i=1}^{n}[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2\widehat{\boldsymbol{x}}_i^T\boldsymbol{\phi}_j(\widehat{\boldsymbol{\beta}},\eta)\widehat{\boldsymbol{x}}_{i,-j*}\Big\|_\infty\widehat{\tau}_j^{-2}\\
\leqslant&\eta\widehat{\tau}_j^{-2}\leqslant\eta\big(\tau_j^{-2}+c_0\sqrt{s_j}\eta\big)=c_1\eta,
\end{aligned}$$

uniformly in j, with probability at least $1-\exp(-c_2\log p)$. By (13) and (14) of the main paper, we have $\Big|1-n^{-1}\sum_{i=1}^{n}[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2\widehat{\boldsymbol{x}}_{i,-1}^T\widehat{\boldsymbol{\theta}}_j\widehat{x}_{i,j}\Big|=|1-\widehat{\tau}_j^{-2}*\widehat{\tau}_j^2|=0$. Hence $||\boldsymbol{I}_{p-1}-\widehat{\boldsymbol{\Theta}}^T\boldsymbol{J}_1||_\infty\leqslant d_0\eta$ with probability at least $1-\exp(-d_1\log p)$, for some positive constants d_0, d_1 and all n sufficiently large. Therefore, there exist some positive constants $c_0\geqslant d_0$, c_1, c_2, such that for all n sufficiently large,

$$\begin{aligned}
P(||\boldsymbol{A}_{n2}||_\infty\geqslant c_0c_2\sqrt{n}s\lambda\eta)\leqslant&P(||\boldsymbol{I}_{p-1}-\widehat{\boldsymbol{\Theta}}^T\boldsymbol{J}_1||_\infty\geqslant c_0\eta)+P(||\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0||_1\geqslant c_2s\lambda)\\
\leqslant&\exp(-c_1\log p),
\end{aligned}$$

where the second inequality applies the above result and Theorem 1. By the assumptions of Theorem 2, we have $\sqrt{n}s\lambda\eta=d_0sh^3\sqrt{n}$ for some positive constant d_0. It completes the proof for (2).

To prove (3), we have

$$
\begin{aligned}
&\boldsymbol{A}_{n3}\\
=&-\sqrt{n}\widehat{\boldsymbol{\Theta}}^T\big[\boldsymbol{S}_n(\widehat{\boldsymbol{\beta}},\widehat{G},\widehat{\mathrm{E}})-\boldsymbol{S}_n(\boldsymbol{\beta}_0,G,\mathrm{E})-\boldsymbol{J}_1(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\big]\\
=&n^{-1/2}\widehat{\boldsymbol{\Theta}}^T\sum_{i=1}^n\widetilde{\epsilon}_iG^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)(\widehat{\boldsymbol{x}}_{i,-1}-\widetilde{\boldsymbol{x}}_{i,-1})\\
&+n^{-1/2}\widehat{\boldsymbol{\Theta}}^T\sum_{i=1}^n\widetilde{\epsilon}_i\big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})-G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\big]\widehat{\boldsymbol{x}}_{i,-1}\\
&+n^{-1/2}\widehat{\boldsymbol{\Theta}}^T\sum_{i=1}^n\big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)-\widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})-\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{x}}_{i,-1}^T(\boldsymbol{\beta}_{0,-1}-\widehat{\boldsymbol{\beta}}_{-1})\big]\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{x}}_{i,-1}\\
\triangleq&\sum_{k=1}^3\boldsymbol{A}_{n3k},
\end{aligned}
$$

where the definition of $\boldsymbol{A}_{n3k}$ $(k=1,2,3)$ is clear from the context. We observe that

$$
\begin{aligned}
&\boldsymbol{A}_{n31}\\
=&n^{-1/2}\boldsymbol{\Theta}^T\sum_{i=1}^n\widetilde{\epsilon}_iG^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)(\widehat{\boldsymbol{x}}_{i,-1}-\widetilde{\boldsymbol{x}}_{i,-1})+n^{-1/2}(\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta})^T\sum_{i=1}^n\widetilde{\epsilon}_iG^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)(\widehat{\boldsymbol{x}}_{i,-1}-\widetilde{\boldsymbol{x}}_{i,-1})\\
\triangleq&\boldsymbol{A}_{n311}+\boldsymbol{A}_{n312},
\end{aligned}
$$

where the definition of $\boldsymbol{A}_{n311}$ and $\boldsymbol{A}_{n312}$ is clear from the context. Lemma A10 implies that $P(||\boldsymbol{A}_{n311}||_\infty\geqslant c_0h\sqrt{s\log(p\vee n)})\leqslant\exp(-c_1\log p)$, for some positive constants c_0, c_1 and all n sufficiently large. Furthermore,

$$
\begin{aligned}
&P\left(||\boldsymbol{A}_{n312}||_\infty\geqslant c_0\widetilde{s}\eta h\sqrt{s\log(p\vee n)}\right)\\
\leqslant&P\left(\Big|\Big|n^{-1/2}\sum_{i=1}^n\widetilde{\epsilon}_iG^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)(\widehat{\boldsymbol{x}}_{i,-1}-\widetilde{\boldsymbol{x}}_{i,-1})\Big|\Big|_\infty\geqslant c_0h\sqrt{s\log(p\vee n)}\right)+P\left(\max_{2\leqslant j\leqslant p}||\widehat{\boldsymbol{\theta}}_j-\boldsymbol{\theta}_j||_1\geqslant\widetilde{s}\eta\right)\\
\leqslant&\exp(-c_1\log p),
\end{aligned}
$$

for some positive constants c_0 and c_1, and all n sufficiently large. In the above, the first inequality applies Hölder's inequality; the second inequality applies lemma A10 and Lemma 2. The assumptions of Theorem 1 and Lemma 2 imply that $h\sqrt{s\log(p\vee n)} \leqslant d_0\sqrt{nh^7} \leqslant d_0 sh^3\sqrt{n}$, and $\widetilde{s}\eta \leqslant d_1$ for some positive constants d_0 and d_1. We have $P(||\boldsymbol{A}_{n31}||_\infty \geqslant c_0\ sh^3\sqrt{n}) \leqslant \exp(-c_1\log p)$, for some positive constants c_0, c_1, and all n sufficiently large.

To verify the bound for $\boldsymbol{A}_{n32}$, we rewrite it as

$$
\begin{aligned}
\boldsymbol{A}_{n32} =& n^{-1/2}\boldsymbol{\Theta}^T\sum_{i=1}^n \widetilde{\epsilon}_i\big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}}) - G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\big]\widetilde{\boldsymbol{x}}_{i,-1} \\
&+ n^{-1/2}(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})^T\sum_{i=1}^n \widetilde{\epsilon}_i\big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}}) - G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\big]\widetilde{\boldsymbol{x}}_{i,-1} \\
&+ n^{-1/2}\widehat{\boldsymbol{\Theta}}^T\sum_{i=1}^n \widetilde{\epsilon}_i\big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}}) - G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\big](\widehat{\boldsymbol{x}}_{i,-1} - \widetilde{\boldsymbol{x}}_{i,-1}) \\
\triangleq& \boldsymbol{A}_{n321} + \boldsymbol{A}_{n322} + \boldsymbol{A}_{n323},
\end{aligned}
$$

with the definition of $\boldsymbol{A}_{n32k}$ $(k = 1, 2, 3)$ is clear from the context. Similarly as the proof for $\boldsymbol{A}_{n311}$ and $\boldsymbol{A}_{n312}$, Lemma A8 implies that $P\left(||\boldsymbol{A}_{n321}||_\infty \geqslant c_0\left[h^2\log(p\vee n)\right]^{1/4}\right) \leqslant \exp(-c_1\log p)$, and $P\left(||\boldsymbol{A}_{n322}||_\infty \geqslant c_0\widetilde{s}\eta\left[h^2\log(p\vee n)\right]^{1/4}\right) \leqslant \exp(-c_1\log p)$, for some positive constants c_0, c_1, and all n sufficiently large. For $\boldsymbol{A}_{n323}$, Lemma B1 implies that for sub-Gaussian random variables $\widetilde{\epsilon}_i$, we have

$$
P\left(\left|n^{-1}\sum_{i=1}^n \widetilde{\epsilon}_i^2 - \mathrm{E}(\widetilde{\epsilon}_i^2)\right| \geqslant 4\sqrt{\sigma_\epsilon^2 + M^2}\sqrt{\frac{\log p}{n}}\right) \leqslant \exp(-c\log p),
$$

for some positive constant c and all n sufficiently large. It implies that $P\Big(n^{-1}\sum_{i=1}^n|\widetilde{\epsilon}_i|^2 \geqslant c_0\Big) \leqslant \exp(-c_1\log p)$, for some positive constants c_0, c_1, and all n sufficiently large. In addition,

$$
\begin{aligned}
&P\Big(n^{-1}\sum_{i=1}^{n}|(\widehat{\boldsymbol{x}}_{i,-1}-\widetilde{\boldsymbol{x}}_{i,-1})^T\widehat{\boldsymbol{\theta}}_j|^2 \geqslant c_2 s h^4\Big)\\
\leqslant & P\Big(n^{-1}\sum_{i=1}^{n}|(\widehat{\boldsymbol{x}}_{i,-1}-\widetilde{\boldsymbol{x}}_{i,-1})^T\boldsymbol{\theta}_j|^2 \geqslant c_2 s h^4/2\Big) + P\Big(n^{-1}\sum_{i=1}^{n}|(\widehat{\boldsymbol{x}}_{i,-1}-\widetilde{\boldsymbol{x}}_{i,-1})^T(\widehat{\boldsymbol{\theta}}_j-\boldsymbol{\theta}_j)|^2 \geqslant c_2 s h^4/2\Big)\\
\leqslant & \exp(-c_3\log p), + P\Big(\Big|\Big|n^{-1}\sum_{i=1}^{n}(\widehat{\boldsymbol{x}}_{i,-1}-\widetilde{\boldsymbol{x}}_{i,-1})^T(\widehat{\boldsymbol{x}}_{i,-1}-\widetilde{\boldsymbol{x}}_{i,-1})\Big|\Big|_\infty ||\widehat{\boldsymbol{\theta}}_j-\boldsymbol{\theta}_j||_1^2 \geqslant c_2 s h^4/2\Big)\\
\leqslant & \exp(-c_4\log p),
\end{aligned}
$$

for some positive constants c_2, c_3, c_4, and all n sufficiently large. In the above, the second inequality applies Lemma A12; the last inequality applies Lemma 2 and Lemma A12. Hence There exist some positive constants d_0, d_1, d_2, d_3, such that for all n sufficiently large,

$$
\begin{aligned}
&P\Big(||\boldsymbol{A}_{n323}||_\infty \geqslant d_0 d_2 h^3\sqrt{sn}\Big)\\
\leqslant & \sum_{j=2}^{p} P\left(\Big|n^{-1}\sum_{i=1}^{n}\widetilde{\epsilon}_i\big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}}) - G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\big](\widehat{\boldsymbol{x}}_{i,-1}-\widetilde{\boldsymbol{x}}_{i,-1})^T\widehat{\boldsymbol{\theta}}_j\Big| \geqslant d_0 d_2 h^3\sqrt{s}\right)\\
\leqslant & P\left(\max_{1\leqslant i\leqslant n}\big|\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}}) - G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\big| \geqslant d_2 h\right)\\
& + \sum_{j=2}^{p} P\left(n^{-1}\sum_{i=1}^{n}|\widetilde{\epsilon}_i| * |(\widehat{\boldsymbol{x}}_{i,-1}-\widetilde{\boldsymbol{x}}_{i,-1})^T\widehat{\boldsymbol{\theta}}_j| \geqslant d_0 h^2\sqrt{s}\right)\\
\leqslant & \exp(-d_1\log p) + P\Big(n^{-1}\sum_{i=1}^{n}|\widetilde{\epsilon}_i|^2 \geqslant d_0\Big) + \sum_{j=2}^{p} P\left(n^{-1}\sum_{i=1}^{n}|(\widehat{\boldsymbol{x}}_{i,-1}-\widetilde{\boldsymbol{x}}_{i,-1})^T\widehat{\boldsymbol{\theta}}_j|^2 \geqslant d_0 s h^4\right)\\
\leqslant & \exp(-d_3\log p),
\end{aligned}
$$

according to Lemma A11, where the second last inequality applies the Cauchy-Schwartz inequality. Note that the assumptions of Theorem 1 and Lemma 2 imply that $\left[h^2\log(p\vee n)\right]^{1/4} \leqslant \sqrt{h\log(p\vee n)} \leqslant d_2\sqrt{h * nh^5} = d_2 h^3\sqrt{n}$, and $\widetilde{s}\eta \leqslant d_3$ for some positive constants d_2 and d_3. Hence there exist positive constants c_0, c_1 such that for all n sufficiently large, we have $P\Big(||\boldsymbol{A}_{n32}||_\infty \geqslant c_0 h^3\sqrt{sn}\Big) \leqslant \exp(-c_1\log p)$.

Finally, let's examine $\boldsymbol{A}_{n33}$. Rewrite it as follows:

$$
\begin{aligned}
&\boldsymbol{A}_{n33} \\
=& n^{-1/2}\widehat{\boldsymbol{\Theta}}^T \sum_{i=1}^{n} \big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) - \widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}}) - \widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{x}}_{i,-1}^T(\boldsymbol{\beta}_{0,-1} - \widehat{\boldsymbol{\beta}}_{-1})\big]\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{x}}_{i,-1} \\
=& n^{-1/2}\boldsymbol{\Theta}^T \sum_{i=1}^{n} \big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) - G(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}}) - G^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{x}}_{i,-1}^T(\boldsymbol{\beta}_{0,-1} - \widehat{\boldsymbol{\beta}}_{-1})\big]\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{x}}_{i,-1} \\
&+ n^{-1/2}(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})^T \sum_{i=1}^{n} \big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) - G(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}}) - G^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{x}}_{i,-1}^T(\boldsymbol{\beta}_{0,-1} - \widehat{\boldsymbol{\beta}}_{-1})\big]\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{x}}_{i,-1} \\
&+ n^{-1/2}\boldsymbol{\Theta}^T \sum_{i=1}^{n} \big[G(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}}) - \widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\big]\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{x}}_{i,-1} \\
&+ n^{-1/2}(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})^T \sum_{i=1}^{n} \big[G(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}}) - \widehat{G}(\widehat{\boldsymbol{x}}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\big]\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{x}}_{i,-1} \\
&+ n^{-1/2}\widehat{\boldsymbol{\Theta}}^T \sum_{i=1}^{n} \big[G^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}}) - \widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\big]\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{x}}_{i,-1}\widehat{\boldsymbol{x}}_{i,-1}^T(\boldsymbol{\beta}_{0,-1} - \widehat{\boldsymbol{\beta}}_{-1}) \\
\triangleq& \sum_{k=1}^{5} \boldsymbol{A}_{n33k},
\end{aligned}
$$

where the definitions of $\boldsymbol{A}_{n33k}$, $k = 1, \cdots, 5$ are clear from the context. Lemma A9 and Lemma 2-(3) together imply that $P(||\boldsymbol{A}_{n331}||_\infty \geqslant c_0 s h^3\sqrt{n}) \leqslant \exp(-c_1 \log p)$, $P(||\boldsymbol{A}_{n332}||_\infty \geqslant c_0\widetilde{s}\eta s h^3\sqrt{n}) \leqslant \exp(-c_1 \log p)$, $P(||\boldsymbol{A}_{n333}||_\infty \geqslant c_0 s h^3\sqrt{n}) \leqslant \exp(-c_1 \log p)$, and $P(||\boldsymbol{A}_{n334}||_\infty \geqslant c_0\widetilde{s}\eta s h^3\sqrt{n}) \leqslant \exp(-c_1 \log p)$, for some positive constants c_0, c_1, and all n sufficiently large. The assumptions of Lemma 2 imply that $\widetilde{s}\eta \leqslant d_0$ for some positive constant d_0. To bound $||\boldsymbol{A}_{n335}||_\infty$, we have

$$
\begin{aligned}
& P\Big(||\boldsymbol{A}_{n335}||_\infty \geqslant c_0\sqrt{n}hs\lambda\Big) \\
\leqslant& P\left(\max_{2\leqslant j\leqslant p}\left\|n^{-1/2}\sum_{i=1}^{n}\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{x}}_{i,-1}\widehat{\boldsymbol{x}}_{i,-1}^T\widehat{\boldsymbol{\theta}}_j\right\|_\infty \geqslant c_0^{1/3}\sqrt{n}\right) \\
&+ P\Big(\max_{1\leqslant i\leqslant n}\big|G^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}}) - \widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\big| \geqslant c_0^{1/3}h\Big) + P\Big(||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_1 \geqslant c_0^{1/3}s\lambda\Big) \\
\leqslant& \exp(-c_1\log p),
\end{aligned}
$$

for some positive constants c_0, c_1, and all n sufficiently large. In the above, the first part of the second inequality applies the proof of Lemma B10; the second and third parts apply Lemma A6 and Theorem 1, respectively. The assumptions of Theorem 1 imply that $\sqrt{n}hs\lambda \leqslant d_0\sqrt{n}sh^3$, for some positive constant d_0. Combining the above results, we conclude that

$$P\Big(||\boldsymbol{A}_{n3}||_\infty \geqslant c_0\Delta_{n,p}\Big) \leqslant \exp(-c_1 \log p),$$

for some positive constants c_0, c_1, and all n sufficiently large. Hence, the theorem is proved.

$\square$

Proof of Corollary 1. Let $\sigma_j^2 = \mathrm{E}\big\{\big[\widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\big]^2\big\}$, which is the $(j-1)^{th}$ diagonal entry of $\boldsymbol{\Theta}^T\boldsymbol{\Lambda}\boldsymbol{\Theta}$. Recall that $\widetilde{\epsilon}_i = 2(2A_i-1)\big[\epsilon_i+g(\boldsymbol{x}_i)\big]$, where ϵ_i is sub-Gaussian independent of $\boldsymbol{x}_i$, and $P\big(\max_{1\leqslant i\leqslant n}|g(\boldsymbol{x}_i)| \leqslant M\big) = 1$. Note that

$$\begin{aligned}
\sigma_j^2 =& \mathrm{E}\big\{\epsilon_i^2\big[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\big]^2\big\} + \mathrm{E}\big\{[g(\boldsymbol{x}_i)]^2\big[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\big]^2\big\}\\
\leqslant& b^2\mathrm{E}(\epsilon_i^2)\mathrm{E}[(\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j)^2] + M^2b^2\mathrm{E}[(\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j)^2]\\
\leqslant& b^2(\sigma_\epsilon^2 + M^2)\xi_1||\boldsymbol{\theta}_j||_2^2,
\end{aligned}$$

where $\max_{1\leqslant i\leqslant n}\big|G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\big| \leqslant b$ by Assumption (A1)-(b), and $\lambda_{\max}(\mathrm{E}[\mathrm{Cov}(\boldsymbol{x}_{-1}|\boldsymbol{x}^T\boldsymbol{\beta}_0)]) \leqslant \xi_1$ by Assumption (A2)-(a). We have

$$\sigma_j^2 \geqslant \mathrm{E}\big\{\epsilon_i^2\big[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\big]^2\big\} \geqslant \mathrm{E}(\epsilon_i^2)\xi_2 \min_{2\leqslant j\leqslant p}||\boldsymbol{\theta}_j||_2^2 \geqslant \sigma_\epsilon^2\xi_2 \min_{2\leqslant j\leqslant p}\tau_{0j}^{-2} \geqslant \sigma_\epsilon^2\xi_2(b^2\xi_1)^{-2},$$

by the smallest eigenvalue condition for $\boldsymbol{\Omega}$ in Assumption (A2)-(a) and Lemma A13. Hence we have

$$\sigma_\epsilon^2\xi_2(b^2\xi_1)^{-2} \leqslant \sigma_j^2 \leqslant (\sigma_\epsilon^2 + M^2)b^2\xi_1||\boldsymbol{\theta}_j||_2^2 \leqslant (\sigma_\epsilon^2 + M^2)b^2\xi_1\xi_2^{-2}.$$

This suggests that $\max_{2\leqslant j\leqslant p}\sigma_j^{-1} \leqslant b^2\xi_1/(\sigma_\epsilon\sqrt{\xi_2})$. Since $\max_{2\leqslant j\leqslant p}\big|\widehat{\Sigma}_{jj} - \sigma_j^2\big| = o_p(1)$ by

Lemma A14, we have $\max_{2\leqslant j\leqslant p}\left|\widehat{\Sigma}_{jj}^{-1/2}-\sigma_j^{-1}\right|=o_p(1)$.

Note that in the proof of Theorem 2, we already showed that $\max_{2\leqslant j\leqslant p}\Big|\widetilde{\beta}_j-\beta_{0j}-\frac{1}{n}\sum_{i=1}^n\widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\Big|=o_p(n^{-1/2})$. Hence

$$
\begin{aligned}
&\max_{2\leqslant j\leqslant p}\left|\widehat{\Sigma}_{j,j}^{-1/2}(\widetilde{\beta}_j-\beta_{0j})-\sigma_j^{-1}\frac{1}{n}\sum_{i=1}^n\widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\right|\\
\leqslant&\max_{2\leqslant j\leqslant p}\sigma_j^{-1}*\max_{2\leqslant j\leqslant p}\left|\widetilde{\beta}_j-\beta_{0j}-\frac{1}{n}\sum_{i=1}^n\widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\right|\\
&+\max_{2\leqslant j\leqslant p}\left|\widehat{\Sigma}_{jj}^{-1/2}-\sigma_j^{-1}\right|*\max_{2\leqslant j\leqslant p}\left|\widetilde{\beta}_j-\beta_{0j}\right|\\
=&o_p(n^{-1/2})*O(1)+o_p(1)*O_p(n^{-1/2})=o_p(n^{-1/2}).
\end{aligned}
$$

Applying the Berry-Esseen bound for CLT, there exists some universal constant $c_0>0$ such that

$$
\begin{aligned}
&\max_{2\leqslant j\leqslant p}\sup_{\alpha\in(0,1)}\left|P\left(\sqrt{n}\left|\sigma_j^{-1}\frac{1}{n}\sum_{i=1}^n\widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\right|\leqslant\Phi^{-1}(1-\alpha/2)\right)-(1-\alpha)\right|\\
\leqslant&\max_{2\leqslant j\leqslant p}\frac{c_0}{\sqrt{n}}\mathrm{E}\left[\left|\widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\right|^3\right]\leqslant\max_{2\leqslant j\leqslant p}\frac{c_0 b^3}{\sqrt{n}}\mathrm{E}\left(|\widetilde{\epsilon}_i\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j|^3\right),
\end{aligned}
$$

where $\Phi(\cdot)$ is the c.d.f of $N(0,1)$, and $\Phi^{-1}(\cdot)$ is the inverse function of $\Phi(\cdot)$. The above probability is bounded by $\frac{c_1}{\sqrt{n}}\mathrm{E}\big(|\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j|^3\big)$, where c_1 does not dependent on n, p and $\boldsymbol{\beta}_0$. Note that $\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j$ is sub-Gaussian with variance proxy $\sigma_x^2||\boldsymbol{\theta}_j||_2^2$. The property of the sub-Gaussian distribution and Lemma A13 imply that $\max_{2\leqslant j\leqslant p}\mathrm{E}\big(|\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j|^3\big)\leqslant c_2||\boldsymbol{\theta}_j||_2^3\leqslant c_2\xi_2^{-3}$, where $c_2>0$ does not depend on n, p and $\boldsymbol{\beta}_0$. Then for a universal constant $c>0$,

$$
\begin{aligned}
&\max_{2\leqslant j\leqslant p}\sup_{\alpha\in(0,1)}\left|P\left(\sqrt{n}\left|\sigma_j^{-1}\frac{1}{n}\sum_{i=1}^n\widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\right|\leqslant\Phi^{-1}(1-\alpha/2)\right)-(1-\alpha)\right|\\
\leqslant&\frac{c}{\sqrt{n}}=o(1).
\end{aligned}
$$

Combining all the results above, we conclude the proof of the corollary. $\square$

Proof of Theorem 3. Let $\widetilde{\delta}_j = n^{-1}\sum_{i=1}^n \widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j$, and $\boldsymbol{\xi} = (\xi_2, \cdots, \xi_p)^T$ be a multivariate mean zero Gaussian with covariance matrix $\boldsymbol{\Theta}^T\boldsymbol{\Lambda}\boldsymbol{\Theta}$.

Since $\widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)$ and $\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j$ are both sub-Gaussian, Comment 2.2 in Chakraborty et al. [2014] implies that $\widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j$ satisfies their condition (E.1) with $B_n = C_1$ for some universal constant C_1, which does not depend on n, p and $\boldsymbol{\beta}_0$. The order of h implies that $\log p = o\big((\sqrt{n}h^3)^{-1}\big)$, and $(nh^5)^{-1} = o(1)$. Hence we can derive that $\frac{[\log(pn)]^7}{n} = O\Big(\frac{[\log p]^7}{n}\Big) = o\big((n^{9/2}h^{21})^{-1}\big) = o(n^{-0.3})$. Note that $\max_{2\leqslant j\leqslant p}|\widetilde{\delta}_j| = \max_{2\leqslant j\leqslant p}\Big\{\widetilde{\delta}_j, -\widetilde{\delta}_j\Big\}$, hence Corollary 2.1 in Chernozhukov et al. [2013] indicates that $\sup_{t\in\mathbb{R}}\big|P(\sqrt{n}\max_{2\leqslant j\leqslant p}|\widetilde{\delta}_j| \leqslant t) - P(\max_{2\leqslant j\leqslant p}|\xi_j| \leqslant t)\big| \leqslant \exp(-c_1\log n)$, for some universal constant $c_1 > 0$.

Define the event $\mathbb{T}_n(\mathcal{G}) = \Big\{\sqrt{n}\Big|\max_{j\in\mathcal{G}}\Big|\widetilde{\beta}_j - \beta_{0j}\Big| - \max_{j\in\mathcal{G}}\Big|\widetilde{\delta}_j\Big|\Big| > \Delta_{n,p}\Big\}$. Note that Theorem 2 implies for some universal constant $c_1 > 0$, and all n sufficiently large,

$$P\big(\mathbb{T}_n(\mathcal{G})\big) \leqslant P\left(\sqrt{n}\max_{j\in\mathcal{G}}\Big|(\widetilde{\beta}_j - \beta_{0j}) - \widetilde{\delta}_j\Big| > \Delta_{n,p}\right) \leqslant P\Big(||\boldsymbol{\Delta}||_\infty \geqslant \Delta_{n,p}\Big) \leqslant \exp(-c_1\log p).$$

Applying Corollary 16 in Wasserman [2014], we have that for some universal positive constants C, c_1, and all n sufficiently large,

$$\begin{aligned}
&P\left(\sqrt{n}\max_{j\in\mathcal{G}}\Big|\widetilde{\beta}_j - \beta_{0j}\Big| \leqslant c^*_{1-\alpha}(\mathcal{G})\right)\\
\leqslant &P\left(\max_{j\in\mathcal{G}}\Big|\widetilde{\delta}_j\Big| \leqslant c^*_{1-\alpha}(\mathcal{G}) + \Delta_{n,p}\right) + P\big(\mathbb{T}_n(\mathcal{G})\big)\\
\leqslant &P\left(\max_{j\in\mathcal{G}}\Big|\widetilde{\delta}_j\Big| \leqslant c^*_{1-\alpha}(\mathcal{G})\right) + C\Delta_{n,p}\sqrt{1 \vee \log(p/\Delta_{n,p})} + \exp(-c_1\log p),
\end{aligned}$$

uniformly over $\alpha \in (0,1)$. Note that $\Delta_{n,p}\sqrt{\log p} = o(1)$ implies that $\Delta_{n,p}\sqrt{1 \vee \log(p/\Delta_{n,p})} = o(1)$. Hence it suffices to show $P\Big(\max_{j\in\mathcal{G}}\Big|\widetilde{\delta}_j\Big| \leqslant c^*_{1-\alpha}\Big) \leqslant 1-\alpha+o(1)$ uniformly over $\alpha \in (0,1)$.

Note that $\max_{j\in\mathcal{G}}|\widetilde{\delta}_j| = \max_{j\in\mathcal{G}}\Big\{\widetilde{\delta}_j, -\widetilde{\delta}_j\Big\}$, and $\max_{j\in\mathcal{G}}|\delta^*_j| = \max_{j\in\mathcal{G}}\big\{\delta^*_j, -\delta^*_j\big\}$. Observe that conditional on $\boldsymbol{w}$, $(\delta^*_1, \cdots, \delta^*_p, -\delta^*_1, \cdots, -\delta^*_p)^T$ is multivariate mean zero Gaussian with covariance matrix $\begin{pmatrix} \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}) & -\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}) \\ -\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}) & \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}) \end{pmatrix}$. Then let $\Delta_0 = ||\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\Theta}^T \boldsymbol{\Lambda} \boldsymbol{\Theta}||_\infty$, Gaussian comparison inequality suggests that for some positive constant C,

$$
\begin{aligned}
P\left(\max_{j\in\mathcal{G}} \left|\widetilde{\delta}_j\right| \leqslant c^*_{1-\alpha}(\mathcal{G})\right) &\leqslant P\left(\max_{j\in\mathcal{G}} \left|\delta^*_j\right| \leqslant c^*_{1-\alpha}(\mathcal{G}) \mid \boldsymbol{w}\right) + C\Delta_0^{1/3}\big[1 \vee \log(p/\Delta_0)\big]^{2/3} \\
&= 1-\alpha + C\Delta_0^{1/3}\big[1 \vee \log(p/\Delta_0)\big]^{2/3},
\end{aligned}
$$

uniformly over $\alpha \in (0,1)$. Proof of Lemma A14 implies that $\Delta_0 \leqslant c_0 \widetilde{s}^{1/2} h$ with probability at least $1-\exp(-c_1 \log p)$, for universal positive constants c_0 and c_1. Hence, we can derive $\Delta_0 \log^2 p = o_p(1)$. It implies that $\Delta_0^{1/3}\big[1 \vee \log(p/\Delta_0)\big]^{2/3} = o(1)$ for all sufficiently large n. We obtain

$$
\sup_{\alpha\in(0,1)} \left[P\Big(\sqrt{n} \max_{j\in\mathcal{G}} \left|\widetilde{\beta}_j - \beta_{0j}\right| \leqslant c^*_{1-\alpha}(\mathcal{G})\Big) - (1-\alpha)\right] = o(1),
$$

for all n sufficiently large. Similarly, we can derive that $\sup_{\alpha\in(0,1)}\Big[(1-\alpha) - P\Big(\sqrt{n}\max_{j\in\mathcal{G}} |\widetilde{\beta}_j - \beta_{0j}| \leqslant c^*_{1-\alpha}(\mathcal{G})\Big)\Big] = o(1)$, for all n sufficiently large. Note that all the universal constants do not depend on n, p and $\boldsymbol{\beta}_0$. We thus have

$$
\sup_{\boldsymbol{\beta}_0\in\mathbb{B}_0:||\boldsymbol{\beta}_0||_0\leqslant s} \sup_{\alpha\in(0,1)} \left|P\Big(\sqrt{n}\max_{j\in\mathcal{G}} |\widetilde{\beta}_j - \beta_{0j}| \leqslant c^*_{1-\alpha}(\mathcal{G})\Big) - (1-\alpha)\right| = o(1).
$$

□

S6 Derivation of the Results in Section S3

Proof of Lemma A1. By the model setup, we have:

$$
\begin{aligned}
&\boldsymbol{S}_n(\boldsymbol{\beta}_0, \widehat{G}, \widehat{\mathrm{E}}) \\
=& -n^{-1}\sum_{i=1}^n \left\{\widetilde{\epsilon}_i + G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) - \widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right\}\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)[\boldsymbol{x}_{i,-1} - \widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)] \\
=& -n^{-1}\sum_{i=1}^n \widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)] \\
& - n^{-1}\sum_{i=1}^n \widetilde{\epsilon}_i \left\{\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) - G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right\}[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)] \\
& - n^{-1}\sum_{i=1}^n \widetilde{\epsilon}_i \widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)[\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - \widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)] \\
& - n^{-1}\sum_{i=1}^n \left\{G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) - \widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right\} G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_{i,-1}^T\boldsymbol{\beta}_0)] \\
& - n^{-1}\sum_{i=1}^n \left\{G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) - \widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right\}\left\{\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) - G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right\}[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)] \\
& - n^{-1}\sum_{i=1}^n \left\{G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) - \widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right\}\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)[\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - \widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)] \\
\triangleq& \sum_{j=1}^6 \boldsymbol{I}_{nj},
\end{aligned}
$$

where the definition of $\boldsymbol{I}_{nj}$ is clear from the context. Since $G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)$ is bounded by Assumption (A1)-(b), and $\widetilde{\epsilon}_i$ is sub-Gaussian by Lemma A2, then $G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\widetilde{\epsilon}_i$ is also sub-Gaussian by Lemma A2. Lemma B1 in Section S7 implies that there exist positive constants c_0, c_1 and c_2 such that for all n sufficiently large,

$$P\left(||\boldsymbol{I}_{n1}||_\infty \geqslant c_0\sqrt{\frac{\log(p\vee n)}{n}}\right) \leqslant \exp[-c_1\log(p\vee n)].$$

Note that $\boldsymbol{\beta}_0 \in \mathbb{B}_1$. Lemma A6 implies that $\max_{1\leqslant i\leqslant n}\left|\widehat{G}^{(1)}(\boldsymbol{x}^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) - G^{(1)}(\boldsymbol{x}^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right| \leqslant c_0 h$, with probability at least $1-\exp[-c_1\log(p\vee n)]$, for some positive constants c_0, c_1, and all

n sufficiently large. Then we can apply the proof of Lemma A8 to show that with probability at least $1-\exp[-c_1\log(p\vee n)]$, $\sqrt{n}||\boldsymbol{I}_{n2}||_\infty \leqslant c_0\left[h^2\log(p\vee n)\right]^{1/4} \leqslant c_0\sqrt{\log(p\vee n)}$. Hence we have that with probability at least $1-\exp[-c_1\log(p\vee n)]$, $||\boldsymbol{I}_{n2}||_\infty \leqslant c_0\sqrt{\frac{\log(p\vee n)}{n}}$.

To bound $||\boldsymbol{I}_{n3}||_\infty$, observe that

$$\begin{aligned}||\boldsymbol{I}_{n3}||_\infty \leqslant&\Big|\Big|n^{-1}\sum_{i=1}^n\widetilde{\epsilon}_iG^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)[\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)-\widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]\Big|\Big|_\infty\\ &+\Big|\Big|n^{-1}\sum_{i=1}^n\widetilde{\epsilon}_i[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)-G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]*[\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)-\widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]\Big|\Big|_\infty\\ =&||\boldsymbol{I}_{n31}||_\infty+||\boldsymbol{I}_{n32}||_\infty,\end{aligned}$$

where the definitions of $\boldsymbol{I}_{n31}$ and $\boldsymbol{I}_{n32}$ are clear from the context. Similarly, Lemma A7 indicates that $\max_{1\leqslant i\leqslant n}\left|\left[\widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\right]^T\boldsymbol{\eta}\right| \leqslant c_0h^2||\boldsymbol{\eta}||_2$, with probability at least $1-\exp[-c_1\log(p\vee n)]$, for some positive constants c_0, c_1, and all n sufficiently large. Then the proof of Lemma A10 implies that $\sqrt{n}||\boldsymbol{I}_{n31}||_\infty \leqslant c_0h\sqrt{s\log(p\vee n)}$ holds with probability at least $1-\exp[-c_1\log(p\vee n)]$, for some positive constants c_0, c_1, and all n sufficiently large. For $||\boldsymbol{I}_{n32}||_\infty$, Lemma A6 and Lemma A7 imply that

$$\begin{aligned}||\boldsymbol{I}_{n32}||_\infty \leqslant& c_0hn^{-1}\sum_{i=1}^n|\widetilde{\epsilon}_i|*\left|\left|\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)-\widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\right|\right|_\infty\\ \leqslant& c_0h^3\sqrt{n^{-1}\sum_{i=1}^n\widetilde{\epsilon}_i^2}\leqslant c_1h^3\leqslant c_0n^{-1/2},\end{aligned}$$

with probability at least $1-\exp[-c_1\log(p\vee n)]$, for some positive constants c_0, c_1, and all n sufficiently large. Hence we have $P\left(||\boldsymbol{I}_{n3}||_\infty \geqslant c_0\sqrt{\frac{\log(p\vee n)}{n}}\right) \leqslant \exp[-c_1\log(p\vee n)]$, for some positive constants c_0, c_1, and all n sufficiently large.

The proof of Lemma A9 implies that $P\left(\sqrt{n}||\boldsymbol{I}_{n4}||_\infty \geqslant h[\log(p\vee n)]^{1/4}\right) \leqslant \exp[-c_1\log(p\vee n)]$. Hence $||\boldsymbol{I}_{n4}||_\infty \leqslant d_0n^{-1/2}$ with probability at least $1-\exp[-c_1\log(p\vee n)]$.

For $\boldsymbol{I}_{n5}$, note that $\boldsymbol{x}_{i,-1}-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)$ is sub-Gaussian. Proof of Lemma B1 indicates that for universal constant $c_1 > 0$,

$$P\left(\frac{1}{n}\sum_{i=1}^{n}\left|\left|[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)][\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^T\right|\right|_\infty \geqslant ||\mathrm{Cov}(\boldsymbol{x}_{-1}|\boldsymbol{x}^T\boldsymbol{\beta}_0)||_\infty + \sigma_x^2\sqrt{\frac{\log(p \vee n)}{n}}\right)$$

$$\leqslant \exp[-c_1 \log(p \vee n)].$$

Lemma A5 and Lemma A6 together indicate $||\boldsymbol{I}_{n5}||_\infty \leqslant ch^3 n^{-1}\sum_{i=1}^n ||\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)||_\infty \leqslant c_0 n^{-1/2}$ with probability at least $1 - \exp[-c_1 \log(p \vee n)]$. Similarly, Lemma A5 and Lemma A7 indicate that $||\boldsymbol{I}_{n6}||_\infty \leqslant c_0 n^{-1/2}$ with the same probability bound. Combining all the previous results, we conclude the lemma. $\square$

Proof of Lemma A2. To prove the first part of the lemma, note that for any unit vector $\boldsymbol{v} \in \mathbb{R}^p$ and $c \in \mathbb{R}$, Jensen's inequality and the sub-Gaussian property of $\boldsymbol{x}$ imply that

$$\begin{aligned}
\mathrm{E}\Big\{\exp[s\boldsymbol{v}^T\mathrm{E}(\boldsymbol{x}\big|\boldsymbol{x}^T\boldsymbol{\beta})]\Big\} &= \mathrm{E}\Big\{\exp[\mathrm{E}(s\boldsymbol{x}^T\boldsymbol{v}\big|\boldsymbol{x}^T\boldsymbol{\beta})]\Big\} \\
&\leqslant \mathrm{E}\Big\{\mathrm{E}[\exp(s\boldsymbol{x}^T\boldsymbol{v})\big|\boldsymbol{x}^T\boldsymbol{\beta}]\Big\} \\
&= \mathrm{E}[\exp(s\boldsymbol{x}^T\boldsymbol{v})] \leqslant \exp\left(\frac{s^2\sigma_x^2}{2}\right).
\end{aligned}$$

For $\boldsymbol{x} - \mathrm{E}(\boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\beta})$, we apply Hölder's inequality, which indicates that for any $q_1, q_2 > 0$ such that $q_1^{-1} + q_2^{-1} = 1$,

$$\begin{aligned}
\mathrm{E}\Big\{\exp\left\{s\boldsymbol{v}^T[\boldsymbol{x} - \mathrm{E}(\boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\beta})]\right\}\Big\} \leqslant& \Big[\mathrm{E}\exp(sq_1\boldsymbol{x}^T\boldsymbol{v})\Big]^{1/q_1}\Big\{\mathrm{E}\exp\left[sq_2\boldsymbol{v}^T\mathrm{E}(\boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\beta})\right]\Big\}^{1/q_2} \\
\leqslant& \mathrm{E}\exp(sq_1\boldsymbol{x}^T\boldsymbol{v}) * \mathrm{E}\exp\left[sq_2\boldsymbol{v}^T\mathrm{E}(\boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\beta})\right] \\
\leqslant& \exp\left[\frac{s^2\sigma_x^2(q_1^2 + q_2^2)}{2}\right].
\end{aligned}$$

Let $q_1 = q_2 = 2$, then we have $\mathrm{E}\Big\{\exp\left\{s\boldsymbol{v}^T[\boldsymbol{x} - \mathrm{E}(\boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\beta})]\right\}\Big\} \leqslant \exp(2s^2\sigma_x^2)$.

For $\widetilde{\epsilon} = 2(2A-1)\big(\epsilon + g(\boldsymbol{x})\big)$, note that ϵ and A are independent. Easy to show that $2(2A-1)\epsilon$ is sub-Gaussian, we note that $\mathrm{E}\big\{\exp[2s(2A-1)\epsilon]\big\} = \frac{1}{2}\big(\mathrm{E}e^{2s\epsilon} + \mathrm{E}e^{-2s\epsilon}\big) \leqslant$

$\exp(2s^2\sigma_\epsilon^2)$. Since $g(\cdot)$ is bounded by M almost everywhere, Exercise 2.4 in Wainwright [2015] implies that $2(2A-1)g(\boldsymbol{x})$ is sub-Gaussian with variance proxy at most $4M^2$. Then similarly as the previous step, we conclude that $\widetilde{\epsilon}$ is sub-Gaussian.

To prove the second part of Lemma A2, denote $z = xy - \mathrm{E}(xy)$. Note that $\mathrm{E}z = 0$. Since y is a sub-Gaussian, for any integer $k \geqslant 1$, we have $\mathrm{E}(|y|^k) \leqslant (2\sigma_y^2)^{k/2} k\Gamma(k/2)$. Hence for any $s > 0$,

$$
\begin{aligned}
&\mathrm{E}[\exp(sz)] \\
\leqslant &1 + \sum_{k=2}^{\infty} \frac{s^k \mathrm{E}(|z|^k)}{k!} \\
\leqslant &1 + \sum_{k=2}^{\infty} \frac{s^k 2^{k-1}\big\{\mathrm{E}(|xy|^k) + [\mathrm{E}(|xy|)]^k\big\}}{k!} \\
\leqslant &1 + \sum_{k=2}^{\infty} \frac{s^k 2^{k-1}\sigma_x^k\big[\mathrm{E}(|y|^k) + (\mathrm{E}|y|)^k\big]}{k!} \\
\leqslant &1 + \sum_{k=2}^{\infty} \frac{(2s\sigma_x)^k \mathrm{E}(|y|^k)}{k!} \\
\leqslant &1 + \sum_{n=1}^{\infty} \frac{(2s\sigma_x)^{2n}(2\sigma_y^2)^n (2n)\Gamma(n)}{(2n)!} + \sum_{n=1}^{\infty} \frac{(2s\sigma_x)^{2n+1}(2\sigma_y^2)^{n+1/2}(2n+1)\Gamma(n+1/2)}{(2n+1)!} \\
\leqslant &1 + (1 + 2\sqrt{2}s\sigma_x\sigma_y) \sum_{n=1}^{\infty} \frac{(2\sqrt{2}s\sigma_x\sigma_y)^{2n} 2(n!)}{(2n)!} \\
\leqslant &1 + (1 + 2\sqrt{2}s\sigma_x\sigma_y) \sum_{n=1}^{\infty} \frac{(2\sqrt{2}s\sigma_x\sigma_y)^{2n}}{n!} \\
= &\exp(8s^2\sigma_x^2\sigma_y^2) + 2\sqrt{2}s\sigma_x\sigma_y\big[\exp(8s^2\sigma_x^2\sigma_y^2) - 1\big] \\
\leqslant &\exp(16s^2\sigma_x^2\sigma_y^2),
\end{aligned}
$$

where the second and the fourth inequalities apply Jensen's inequality, the second last inequality applies that $2(n!)^2 \leqslant (2n)!$ for any $n \geqslant 1$. The conclusion follows by the definition of sub-Gaussian random variables. $\square$

Proof of Lemma A3. Note that $\boldsymbol{x}_i$, $[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})]^T\boldsymbol{\theta}_j$ and $\widetilde{\epsilon}_i$ are all sub-Gaussian with variance proxy σ_x^2, $2||\boldsymbol{\theta}_j||_2^2\sigma_x^2$, and $4(\sigma_\epsilon^2 + M^2)$, respectively, by Lemma A2. Lemma A13 implies that $\max_{2\leqslant j\leqslant p}||\boldsymbol{\theta}_j||_0 \leqslant \widetilde{s}+1$, and $||\boldsymbol{\theta}_j||_2 \leqslant \xi_2^{-1}$ uniformly in j, where ξ_2 is defined in Assumption (A2)-(a). Denote $\breve{\boldsymbol{x}}_{i,-1} = \boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})$, and $\widetilde{\boldsymbol{x}}_{i,-1} = \boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)$. Assumption (A2)-(c) and Lemma B2 imply that

$$
\begin{aligned}
&\sup_{\boldsymbol{\beta}\in\mathbb{B}} n^{-1}\sum_{i=1}^n \left|\breve{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j - \widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\right|^2 \\
\leqslant & C\xi_2^{-2}\sup_{\boldsymbol{\beta}\in\mathbb{B}} n^{-1}\sum_{i=1}^n \left(|\boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{x}_i^T\boldsymbol{\beta}_0| + \max(|\boldsymbol{x}_i^T\boldsymbol{\beta}|, |\boldsymbol{x}_i^T\boldsymbol{\beta}_0|)||\boldsymbol{\beta}-\boldsymbol{\beta}_0||_2\right)^2 \\
\leqslant & 2C\xi_2^{-2}\sup_{\boldsymbol{\beta}\in\mathbb{B}} n^{-1}\sum_{i=1}^n \left(|\boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{x}_i^T\boldsymbol{\beta}_0|^2 + (|\boldsymbol{x}_i^T\boldsymbol{\beta}|^2 + |\boldsymbol{x}_i^T\boldsymbol{\beta}_0|^2)||\boldsymbol{\beta}-\boldsymbol{\beta}_0||_2^2\right) \\
\leqslant & 2C\xi_2^{-2}\sup_{\boldsymbol{\beta}\in\mathbb{B}} n^{-1}\sum_{i=1}^n |\boldsymbol{x}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)|^2 + 4Cr^2\xi_2^{-2}\sup_{\boldsymbol{\beta}\in\mathbb{B}} n^{-1}\sum_{i=1}^n |\boldsymbol{x}_i^T\boldsymbol{\beta}|^2 \\
\leqslant & c_0\xi_2^{-2}\sigma_x^2,
\end{aligned}
$$

with probability at least $1-\exp(-c_1 n)$, for some positive constants C, c_0, c_1 and all n sufficiently large, where the last inequality applies Lemma B2, since the sub-Gaussian property implies that $\mathrm{E}[(\boldsymbol{x}_i^T\boldsymbol{v})^2] \leqslant \sigma_x^2$ for any $||\boldsymbol{v}||_2 = 1$. We thus have

$$
\begin{aligned}
&P\left(\max_{2\leqslant j\leqslant p}\sup_{\boldsymbol{\beta}\in\mathbb{B}} n^{-1}\sum_{i=1}^n \left|\breve{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\right|^2 \geqslant d_0\xi_2^{-2}\sigma_x^2\right) \\
\leqslant & \sum_{j=2}^p P\left(\sup_{\boldsymbol{\beta}\in\mathbb{B}} n^{-1}\sum_{i=1}^n \left|\breve{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\right|^2 \geqslant d_0\xi_2^{-2}\sigma_x^2\right) \\
\leqslant & \sum_{j=2}^p P\left(\sup_{\boldsymbol{\beta}\in\mathbb{B}} 2n^{-1}\sum_{i=1}^n \left|\breve{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j - \widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\right|^2 \geqslant 2c_0\xi_2^{-2}\sigma_x^2\right) + P\left(2n^{-1}\sum_{i=1}^n \left|\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\right|^2 \geqslant (d_0-2c_0)\xi_2^{-2}\sigma_x^2\right) \\
\leqslant & (p-1)\exp(-c_2 n) + P\left(\left|n^{-1}\sum_{i=1}^n \left|\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\right|^2 - \mathrm{E}\left(\left|\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\right|^2\right)\right| \geqslant (d_0-2c_0-4)\xi_2^{-2}\sigma_x^2/2\right) \\
\leqslant & p\exp(-c_2 n) = \exp(-c_2 n + \log p) = \exp(-cn),
\end{aligned}
$$

for some positive constants $d_0 > 4$, c_2, c, and all n sufficiently large. In the above, the last inequality applies Lemma B1, with $\max_{2\leqslant j\leqslant p}\mathrm{E}(|\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j|^2) \leqslant 2\xi_2^{-2}\sigma_x^2$ by its sub-Gaussian property. It derives the probability bound for $\mathcal{G}_n$. The probability bound for $\mathcal{K}_n$ follows from Lemma B2 with similar technique. For $\mathcal{H}_n$, note that $\left|2[\epsilon_i + g(\boldsymbol{x}_i)] * \left[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})\right]^T \boldsymbol{\theta}_j\right|^2 = \left|\widetilde{\epsilon}_i \left[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})\right]^T \boldsymbol{\theta}_j\right|^2$, where $\widetilde{\epsilon}_i = 2(2A_i - 1)[\epsilon_i + g(\boldsymbol{x}_i)]$. Lemma A2 implies that $\breve{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j$ and $\widetilde{\epsilon}_i$ are both sub-Gaussian. Hence we have that $\mathrm{E}(\widetilde{\epsilon}_i^4) \leqslant 16 * \left[4(\sigma_\epsilon^2 + M^2)\right]^2$, and $\max_{2\leqslant j\leqslant p} \mathrm{E}\left[(\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j)^4\right] \leqslant 16 * \left(2\xi_2^{-2}\sigma_x^2\right)^2$. Similar as the above analysis for $\mathcal{G}_n$, Assumption (A2)-(c) implies that

$$
\begin{aligned}
\sup_{\boldsymbol{\beta}\in\mathbb{B}} n^{-1}\sum_{i=1}^n \left|\breve{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j - \widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\right|^4 \leqslant & C\xi_2^{-4}\sup_{\boldsymbol{\beta}\in\mathbb{B}} n^{-1}\sum_{i=1}^n |\boldsymbol{x}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)|^4 + 2Cr^4\xi_2^{-4}\sup_{\boldsymbol{\beta}\in\mathbb{B}} n^{-1}\sum_{i=1}^n |\boldsymbol{x}_i^T\boldsymbol{\beta}|^4 \\
\leqslant & c_0\xi_2^{-4}\sigma_x^4,
\end{aligned}
$$

with probability at least $1 - \exp(-c_1\sqrt{n})$, for some positive constants C, c_0, c_1 and all n sufficiently large, where the last inequality applies Lemma B3 with $s_0 = ks$ and $t = c_3\xi_2^{-4}\sigma_x^4$ for some positive constant c_3. Hence by Lemma B3, there exist some positive constants $d_1 > 256\sqrt{2}$, $d_2 > 16$, c_2, c, such that for all n sufficiently large,

$$
\begin{aligned}
& P\left(\max_{2\leqslant j\leqslant p}\sup_{\boldsymbol{\beta}\in\mathbb{B}} n^{-1}\sum_{i=1}^n \left|\widetilde{\epsilon}_i\breve{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\right|^2 \geqslant d_1\xi_2^{-2}\sigma_x^2(\sigma_\epsilon^2 + M^2)\right) \\
\leqslant & P\left(\sqrt{n^{-1}\sum_{i=1}^n \widetilde{\epsilon}_i^4} \geqslant d_2(\sigma_\epsilon^2 + M^2)\right) + P\left(\max_{2\leqslant j\leqslant p}\sup_{\boldsymbol{\beta}\in\mathbb{B}}\sqrt{n^{-1}\sum_{i=1}^n \left|\breve{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\right|^4} \geqslant d_1 d_2^{-1}\xi_2^{-2}\sigma_x^2\right) \\
\leqslant & P\left(n^{-1}\sum_{i=1}^n \widetilde{\epsilon}_i^4 \geqslant d_2^2(\sigma_\epsilon^2 + M^2)^2\right) + \sum_{j=2}^p P\left(\sup_{\boldsymbol{\beta}\in\mathbb{B}} n^{-1}\sum_{i=1}^n \left|\breve{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\right|^4 \geqslant d_1^2 d_2^{-2}\xi_2^{-4}\sigma_x^4\right) \\
\leqslant & P\left(\left|n^{-1}\sum_{i=1}^n \widetilde{\epsilon}_i^4 - \mathrm{E}\left(\widetilde{\epsilon}_i^4\right)\right| \geqslant (d_2^2 - 256)(\sigma_\epsilon^2 + M^2)^2\right) + P\left(8n^{-1}\sum_{i=1}^n \left(\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\right)^4 \geqslant (d_1^2 d_2^{-2} - 8c_0)\xi_2^{-4}\sigma_x^4\right) \\
& + \sum_{j=2}^p P\left(\sup_{\boldsymbol{\beta}\in\mathbb{B}} 8n^{-1}\sum_{i=1}^n \left(\breve{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j - \widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\right)^4 \geqslant 8c_0\xi_2^{-4}\sigma_x^4\right) \\
\leqslant & p\exp(-c_2\sqrt{n}) + P\left(8\left|n^{-1}\sum_{i=1}^n \left(\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\right)^4 - \mathrm{E}\left[\left(\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\right)^4\right]\right| \geqslant (d_1^2 d_2^{-2} - 512 - 8c_0)\xi_2^{-4}\sigma_x^4\right) \\
\leqslant & (p+1)\exp(-c_2\sqrt{n}) = \exp(-c\sqrt{n}).
\end{aligned}
$$

Hence we prove that $P(\mathcal{H}_n) \geqslant 1 - \exp(-c\sqrt{n})$, for some positive constant c and all n sufficiently large.

For $\mathcal{J}_n$, note that

$$\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}|\boldsymbol{x}_i^T\boldsymbol{\beta}| \leqslant \max_{1\leqslant i\leqslant n}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0| + \max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|\boldsymbol{x}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right|.$$

The sub-Gaussian property of $\boldsymbol{x}_i^T\boldsymbol{\beta}_0$ implies that $P\left(\max_{1\leqslant i\leqslant n}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0| \geqslant ||\boldsymbol{\beta}_0||_2\sigma_x\sqrt{\log(p\vee n)}\right) \leqslant \exp[-c\log(p\vee n)]$, for some positive constant c, and all n sufficiently large. Note that for any $\boldsymbol{\beta}\in\mathbb{B}_1$, we have that $||\boldsymbol{\beta}-\boldsymbol{\beta}_0||_2 \leqslant c_0\sqrt{s}h^2$, and $||\boldsymbol{\beta}-\boldsymbol{\beta}_0||_0 \leqslant (k+1)s$. Then we have that $\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|\boldsymbol{x}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right| \leqslant \max_{1\leqslant i\leqslant n}||\boldsymbol{x}_i||_\infty\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}||\boldsymbol{\beta}-\boldsymbol{\beta}_0||_1 \leqslant c_0 s^{3/2}h^2\sqrt{\log(p\vee n)}$ with probability at least $1-\exp[-c_1\log(p\vee n)]$, for some positive constants c_0, c_1 and all n sufficiently large, according to (S1). The assumptions of Theorem 1 imply that $s^{3/2}h^2 \leqslant 1$. Hence we conclude that

$$P\left(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}|\boldsymbol{x}_i^T\boldsymbol{\beta}| \geqslant 2||\boldsymbol{\beta}_0||_2\sigma_x\sqrt{\log(p\vee n)}\right) \leqslant \exp[-c\log(p\vee n)],$$

for some positive constant c, and all n sufficiently large. It concludes the proof of the lemma. □

Proof of Lemma A4. Recall $\boldsymbol{\gamma} = \boldsymbol{\beta}_{-1} - \boldsymbol{\beta}_{0,-1}$. By Taylor expansion, we can derive that

$$\begin{aligned}
G(t|\boldsymbol{\beta}) =& \mathrm{E}\big[f_0(\boldsymbol{x}^T\boldsymbol{\beta}_0)|\boldsymbol{x}^T\boldsymbol{\beta}=t\big] \\
=& \mathrm{E}\left[f_0(\boldsymbol{x}^T\boldsymbol{\beta}) - f_0'(\boldsymbol{x}^T\boldsymbol{\beta})\boldsymbol{x}_{-1}^T\boldsymbol{\gamma} + \int_{\boldsymbol{x}^T\boldsymbol{\beta}}^{\boldsymbol{x}^T\boldsymbol{\beta}_0} f_0''(u)(\boldsymbol{x}^T\boldsymbol{\beta}_0-u)du|\boldsymbol{x}^T\boldsymbol{\beta}=t\right] \\
=& f_0(t) - f_0'(t)\mathrm{E}(\boldsymbol{x}_{-1}|\boldsymbol{x}^T\boldsymbol{\beta}=t)^T\boldsymbol{\gamma} + \mathrm{E}\left[\int_0^{\boldsymbol{x}_{-1}^T\boldsymbol{\gamma}} a f_0''(a+\boldsymbol{x}^T\boldsymbol{\beta}_0)da|\boldsymbol{x}^T\boldsymbol{\beta}=t\right]. \quad \text{(S25)}
\end{aligned}$$

Hence we have

$$
\begin{aligned}
G(\boldsymbol{x}^T\boldsymbol{\beta}|\boldsymbol{\beta}) - G(\boldsymbol{x}^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) =& \big[f_0(\boldsymbol{x}^T\boldsymbol{\beta}) - f_0(\boldsymbol{x}^T\boldsymbol{\beta}_0)\big] - f_0'(\boldsymbol{x}^T\boldsymbol{\beta})\mathrm{E}(\boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\beta})^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0) \\
&+ \mathrm{E}\left[\int_0^{\boldsymbol{x}^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)} a f_0''(a+\boldsymbol{x}^T\boldsymbol{\beta}_0)da|\boldsymbol{x}^T\boldsymbol{\beta}\right] \\
=& f_0'(\boldsymbol{x}^T\boldsymbol{\beta})\left[\boldsymbol{x}_{-1}^T\boldsymbol{\gamma} - \mathrm{E}(\boldsymbol{x}_{-1}^T\boldsymbol{\gamma}|\boldsymbol{x}^T\boldsymbol{\beta})\right] - \int_0^{\boldsymbol{x}_{-1}^T\boldsymbol{\gamma}} a f_0''(a+\boldsymbol{x}^T\boldsymbol{\beta}_0)da \\
&+ \mathrm{E}\left[\int_0^{\boldsymbol{x}_{-1}^T\boldsymbol{\gamma}} a f_0''(a+\boldsymbol{x}^T\boldsymbol{\beta}_0)da|\boldsymbol{x}^T\boldsymbol{\beta}\right].
\end{aligned}
$$

It proves (S2). To prove (S3), according to equation (S25), we have that

$$
\begin{aligned}
G^{(1)}(t|\boldsymbol{\beta}) = \frac{d}{dt}G(t|\boldsymbol{\beta}) =& f_0'(t) - f_0'(t)\mathrm{E}^{(1)}(\boldsymbol{x}_{-1}|\boldsymbol{x}^T\boldsymbol{\beta}=t)^T\boldsymbol{\gamma} - f_0''(t)\mathrm{E}(\boldsymbol{x}_{-1}|\boldsymbol{x}^T\boldsymbol{\beta}=t)^T\boldsymbol{\gamma} \\
&+ \frac{d}{dt}\mathrm{E}\left[\int_0^{\boldsymbol{x}_{-1}^T\boldsymbol{\gamma}} a f_0''(a+\boldsymbol{x}^T\boldsymbol{\beta}_0)da|\boldsymbol{x}^T\boldsymbol{\beta}=t\right] \\
=& f_0'(t) - f_0'(t)\mathrm{E}^{(1)}(\boldsymbol{x}_{-1}|\boldsymbol{x}^T\boldsymbol{\beta}=t)^T\boldsymbol{\gamma} - f_0''(t)\mathrm{E}(\boldsymbol{x}_{-1}|\boldsymbol{x}^T\boldsymbol{\beta}=t)^T\boldsymbol{\gamma} \\
&+ \mathrm{E}^{(1)}\left[\int_0^{\boldsymbol{x}_{-1}^T\boldsymbol{\gamma}} a f_0''(a+\boldsymbol{x}^T\boldsymbol{\beta}_0)da|\boldsymbol{x}^T\boldsymbol{\beta}=t\right],
\end{aligned}
$$

where $\mathrm{E}^{(1)}(\cdot|\boldsymbol{x}^T\boldsymbol{\beta}=t)$ is the first derivative of $\mathrm{E}(\cdot|\boldsymbol{x}^T\boldsymbol{\beta}=t)$ with respect to t. Hence we have

$$
\begin{aligned}
&G^{(1)}(\boldsymbol{x}^T\boldsymbol{\beta}|\boldsymbol{\beta}) - G^{(1)}(\boldsymbol{x}^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) \\
=& f_0''(\boldsymbol{x}^T\boldsymbol{\beta})\big[\boldsymbol{x}_{-1}^T\boldsymbol{\gamma} - \mathrm{E}(\boldsymbol{x}_{-1}^T\boldsymbol{\gamma}|\boldsymbol{x}^T\boldsymbol{\beta})\big] - \int_0^{\boldsymbol{x}_{-1}^T\boldsymbol{\gamma}} a f_0'''(a+\boldsymbol{x}^T\boldsymbol{\beta}_0)da \\
&- f_0'(\boldsymbol{x}^T\boldsymbol{\beta})\mathrm{E}^{(1)}(\boldsymbol{x}_{-1}|\boldsymbol{x}^T\boldsymbol{\beta}=t)^T\boldsymbol{\gamma} + \mathrm{E}^{(1)}\left[\int_0^{\boldsymbol{x}_{-1}^T\boldsymbol{\gamma}} a f_0''(a+\boldsymbol{x}^T\boldsymbol{\beta}_0)da|\boldsymbol{x}^T\boldsymbol{\beta}=t\right],
\end{aligned}
$$

It proves (S3).

Denote the event

$$\begin{aligned}\mathcal{E}_0 = &\left\{\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}}|\boldsymbol{x}_i^T\boldsymbol{\beta}| \geqslant c_0\sqrt{s\log(p\vee n)}\right\}\\ &\bigcap\left\{\max_{1\leqslant i\leqslant n}|\boldsymbol{x}_i^T(\boldsymbol{\beta}_1-\boldsymbol{\beta}_2)| \geqslant c_0\sqrt{s\log(p\vee n)}||\boldsymbol{\beta}_1-\boldsymbol{\beta}_2||_2,\ \forall\ \boldsymbol{\beta}_1,\boldsymbol{\beta}_2\in\mathbb{B}\right\},\end{aligned}$$

for some constant $c_0 > 0$. Taking $t = d_0\sigma_x^2\log(p\vee n)$ and $s_0 = ks$, Lemma B2 implies that

$$\begin{aligned}P\left(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{v}\in\mathbb{K}(2ks)}(\boldsymbol{x}_i^T\boldsymbol{v})^2 \leqslant (d_0+1)\sigma_x^2 s\log(p\vee n)\right) &\leqslant \sum_{i=1}^n P\left(\sup_{\boldsymbol{v}\in\mathbb{K}(2ks)}(\boldsymbol{x}_i^T\boldsymbol{v})^2 \leqslant (d_0+1)\sigma_x^2 s\log(p\vee n)\right)\\ &\leqslant \exp[-d_1 s\log(p\vee n)],\end{aligned}$$

for some positive constants d_0, d_1, and all n sufficiently large. Combining this result with the definition of $\mathbb{B}$, we obtain that $P(\mathcal{E}_0) \geqslant 1-\exp[-c_2 s\log(p\vee n)]$, for some positive constants c_0, c_2, and all n sufficiently large. Note that $G(t|\boldsymbol{\beta})$ is twice-differentiable with respect to t, and the derivatives are bounded by Assumption (A5)-(a). Hence on the event $\mathcal{E}_0$, we have

$$\begin{aligned}&\max_{1\leqslant i\leqslant n}\left[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_1|\boldsymbol{\beta}_1) - G(\boldsymbol{x}_i^T\boldsymbol{\beta}_2|\boldsymbol{\beta}_2)\right]^2\\ \leqslant &\max_{1\leqslant i\leqslant n} 2\left[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_1|\boldsymbol{\beta}_1) - G(\boldsymbol{x}_i^T\boldsymbol{\beta}_1|\boldsymbol{\beta}_2)\right]^2 + \max_{1\leqslant i\leqslant n} 2\left[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_1|\boldsymbol{\beta}_2) - G(\boldsymbol{x}_i^T\boldsymbol{\beta}_2|\boldsymbol{\beta}_2)\right]^2\\ \leqslant &c_3||\boldsymbol{\beta}_1-\boldsymbol{\beta}_2||_2 s\log(p\vee n) + c_3\max_{1\leqslant i\leqslant n}\left[\boldsymbol{x}_i^T(\boldsymbol{\beta}_1-\boldsymbol{\beta}_2)\right]^2\\ \leqslant &c_1||\boldsymbol{\beta}_1-\boldsymbol{\beta}_2||_2 s\log(p\vee n),\end{aligned}$$

for any $\boldsymbol{\beta}_1,\boldsymbol{\beta}_2\in\mathbb{B}$, some positive constants c_1 and c_3, where the first part of the second inequality applies Assumption (A5)-(c). It proves (S4). We can conclude (S5) with similar techniques.

To prove (S6), observe that

$$n^{-1}\sum_{i=1}^{n}\big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\big]^2 \leqslant 3\sum_{k=1}^{3}A_k,$$

where

$$\begin{aligned}
A_1 &= n^{-1}\sum_{i=1}^{n}\big[f_0'(\boldsymbol{x}_i^T\boldsymbol{\beta})\big]^2\big[\boldsymbol{x}_{i,-1}^T\boldsymbol{\gamma}-\mathrm{E}(\boldsymbol{x}_{i,-1}^T\boldsymbol{\gamma}|\boldsymbol{x}_i^T\boldsymbol{\beta})\big]^2,\\
A_2 &= n^{-1}\sum_{i=1}^{n}\big[h(\boldsymbol{x}_{i,-1}^T\boldsymbol{\gamma})\big]^2,\\
A_3 &= n^{-1}\sum_{i=1}^{n}\big\{\mathrm{E}\big[h(\boldsymbol{x}_{i,-1}^T\boldsymbol{\gamma})|\boldsymbol{x}_i^T\boldsymbol{\beta}\big]\big\}^2,
\end{aligned}$$

with $h(u)=\int_0^u af_0''(a+\boldsymbol{x}^T\boldsymbol{\beta}_0)da$. It is sufficient to bound A_k for $k=1,2,3$. To bound A_1, we have

$$A_1 \leqslant b^2 n^{-1}\sum_{i=1}^{n}\big[\boldsymbol{x}_{i,-1}^T\boldsymbol{\gamma}-\mathrm{E}(\boldsymbol{x}_{i,-1}^T\boldsymbol{\gamma}|\boldsymbol{x}_i^T\boldsymbol{\beta})\big]^2.$$

Lemma B7 in Section S7 implies that $A_1\leqslant c_1||\boldsymbol{\gamma}||_2^2$ with probability at least $1-\exp(-c_2n)$, for positive constants c_1, c_2. For A_2, note that

$$A_2 = (4n)^{-1}\sum_{i=1}^{n}\Big[f_0''(z_i)(\boldsymbol{x}_{i,-1}^T\boldsymbol{\gamma})^2\Big]^2 \leqslant c_0n^{-1}\sum_{i=1}^{n}(\boldsymbol{x}_{i,-1}^T\boldsymbol{\gamma})^4,$$

where z_i is between $\boldsymbol{x}^T\boldsymbol{\beta}$ and $\boldsymbol{x}^T\boldsymbol{\beta}_0$, for positive constant c_0. The last inequality applies the assumption that $f_0''(\cdot)$ is bounded. Lemma B3 indicates that $A_2\leqslant c_1||\boldsymbol{\gamma}||_2^4$ with probability at least $1-\exp(-c_2\sqrt{n})$, since $s\log p\leqslant d_0nh^5\leqslant d_1n^{1/6}\leqslant d_1\sqrt{n}$, for some positive constants d_0, d_1 and $n\geqslant 1$.

For A_3, observe that

$$A_3 = (4n)^{-1}\sum_{i=1}^{n}\big\{\mathrm{E}\big[f_0''(z_i)(\boldsymbol{x}_{i,-1}^T\boldsymbol{\gamma})^2|\boldsymbol{x}_i^T\boldsymbol{\beta}\big]\big\}^2 \leqslant c_0 n^{-1}\sum_{i=1}^{n}\big[\boldsymbol{\gamma}^T\mathrm{E}(\boldsymbol{x}_{i,-1}\boldsymbol{x}_{i,-1}^T|\boldsymbol{x}_i^T\boldsymbol{\beta})\boldsymbol{\gamma}\big]^2,$$

where z_i is between $\boldsymbol{x}^T\boldsymbol{\beta}$ and $\boldsymbol{x}^T\boldsymbol{\beta}_0$. Assumption (A2) implies that

$$n^{-1}\sum_{i=1}^{n}\big[\boldsymbol{\gamma}^T\mathrm{E}(\boldsymbol{x}_{i,-1}\boldsymbol{x}_{i,-1}^T|\boldsymbol{x}_i^T\boldsymbol{\beta})\boldsymbol{\gamma}\big]^2 \leqslant \xi_4||\boldsymbol{\gamma}||_2^4.$$

Hence we obtain the high probability upper bound of A_3. Then combining all the above results, we complete the proof for (S6). □

Proof of Lemma A5. Note that

$$\begin{aligned}
\widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) &= \sum_{i=1}^{n} W_{ni}(t|\boldsymbol{\beta})\big(\widetilde{Y}_i - G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big)\\
&= \frac{(n-1)^{-1}\sum_{j=1,j\neq i}^{n} K_h(\boldsymbol{x}_i^T\boldsymbol{\beta}-\boldsymbol{x}_j^T\boldsymbol{\beta})\big[\widetilde{Y}_j - G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]}{(n-1)^{-1}\sum_{j=1,j\neq i}^{n} K_h(\boldsymbol{x}_i^T\boldsymbol{\beta}-\boldsymbol{x}_j^T\boldsymbol{\beta})}\\
&\triangleq \frac{A_{n1}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) + A_{n2}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})}{A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})},
\end{aligned}$$

where

$$\begin{aligned}
A_{n1}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) =&(n-1)^{-1}\sum_{j=1,j\neq i}^{n} K_h(\boldsymbol{x}_i^T\boldsymbol{\beta}-\boldsymbol{x}_j^T\boldsymbol{\beta})\widetilde{\epsilon}_j,\\
A_{n2}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) =&(n-1)^{-1}\sum_{j=1,j\neq j}^{n} K_h(\boldsymbol{x}_i^T\boldsymbol{\beta}-\boldsymbol{x}_j^T\boldsymbol{\beta})\big[f_0(\boldsymbol{x}_j^T\boldsymbol{\beta}_0) - G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big],\\
A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) =&(n-1)^{-1}\sum_{j=1,j\neq i}^{n} K_h(\boldsymbol{x}_i^T\boldsymbol{\beta}-\boldsymbol{x}_i^T\boldsymbol{\beta}).
\end{aligned}$$

Then Lemma B4–B6 provide the high probability bounds for A_{ni}, $i = 1, 2, 3$, as following:

$$P\left(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}}|A_{n1}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})| \geqslant c_0h^2\right) \leqslant \exp(-c_1nh^5),$$
$$P\left(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}}|A_{n2}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})| \geqslant c_0h^2\right) \leqslant \exp\big[-c_1\log(p\vee n)\big],$$
$$P\left(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}}\big|A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - \mathrm{E}\big[A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]\big| \geqslant c_0h^2\right) \leqslant \exp(-c_1nh^5).$$

for universal positive constants c_0, c_1, and all n sufficiently large.

We denote the p.d.f of $\boldsymbol{x}^T\boldsymbol{\beta}$ as $f_{\boldsymbol{\beta}}(\cdot)$. For $\mathrm{E}\big[A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]$, we have

$$\begin{aligned}
\mathrm{E}\big[A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big] &= h^{-1}\int K\Big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}-y}{h}\Big)f_{\boldsymbol{\beta}}(y)dy \\
&= \int K(-z)f_{\boldsymbol{\beta}}(\boldsymbol{x}_i^T\boldsymbol{\beta}+hz)dz \\
&= \int K(-z)\Big[f_{\boldsymbol{\beta}}(\boldsymbol{x}_i^T\boldsymbol{\beta}) + hzf'_{\boldsymbol{\beta}}(\boldsymbol{x}_i^T\boldsymbol{\beta}) + \frac{h^2z^2}{2}f''_{\boldsymbol{\beta}}(\widetilde{t})\Big]dz \\
&= f_{\boldsymbol{\beta}}(\boldsymbol{x}_i^T\boldsymbol{\beta}) + \frac{h^2}{2}\int z^2K(-z)f''_{\boldsymbol{\beta}}(\widetilde{t})dz,
\end{aligned}$$

where $\widetilde{t}$ is between $\boldsymbol{x}_i^T\boldsymbol{\beta}$ and $\boldsymbol{x}_i^T\boldsymbol{\beta}+hz$. In the above, the second equality employs the transformation $z = (y - \boldsymbol{x}_i^T\boldsymbol{\beta})/h$. Assumption (A3)–(A4) imply that

$$\sup_{\boldsymbol{\beta}\in\mathbb{B}}\left|\frac{h^2}{2}\int z^2K(-z)f''_{\boldsymbol{\beta}}(\widetilde{t})dz\right| \leqslant c_0h^2,$$

for some positive constant c_0. Hence $f_{\boldsymbol{\beta}}(\boldsymbol{x}_i^T\boldsymbol{\beta}) - c_0h^2 \leqslant \mathrm{E}\big[A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big] \leqslant f_{\boldsymbol{\beta}}(\boldsymbol{x}_i^T\boldsymbol{\beta}) + c_0h^2$. Assumption (A4) implies that $\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}} f_{\boldsymbol{\beta}}^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}) \leqslant M$, for some positive constant M. It ensures that

$$P\left(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}}[A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})]^{-1} \geqslant 2M\right) \leqslant \exp(-c_1nh^5), \tag{S26}$$

for universal positive constant c_1, and all n sufficiently large. Then we conclude that for universal positive constants c_0 and c_1, and all n sufficiently large,

$$P\left(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}}\left|\widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\right|\geqslant c_0h^2\right)\leqslant\exp\big[-c_1\log(p\vee n)\big].$$

$\square$

Proof of Lemma A6. Since

$$\widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})=\frac{A_{n1}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})+A_{n2}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})}{A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})},$$

we have

$$\begin{aligned}\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\triangleq&\frac{A_{n1}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})+A_{n2}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})}{A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})}\\&+\frac{A_{n1}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})+A_{n2}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})}{A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})}*\frac{A_{n3}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})}{A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})},\end{aligned}$$

where $G^{(1)}(t|\boldsymbol{\beta})=\frac{d}{dt}G(t|\boldsymbol{\beta})$, $\widehat{G}(t|\boldsymbol{\beta})=\frac{d}{dt}\widehat{G}(t|\boldsymbol{\beta})$, $A_{nk}^{(1)}(t|\boldsymbol{\beta})=\frac{d}{dt}A_{nk}(t|\boldsymbol{\beta})$, for $k=1,2,3$. Let $K_h(z)=h^{-1}K(z/h)$, and $K_h'(z)=h^{-2}K'(z/h)$. We have

$$\begin{aligned}A_{n1}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})=&(n-1)^{-1}\sum_{j=1,j\neq i}^{n}K_h'(\boldsymbol{x}_i^T\boldsymbol{\beta}-\boldsymbol{x}_j^T\boldsymbol{\beta})\widetilde{\epsilon}_j,\\A_{n2}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})=&(n-1)^{-1}\sum_{j=1,j\neq i}^{n}K_h'(\boldsymbol{x}_i^T\boldsymbol{\beta}-\boldsymbol{x}_j^T\boldsymbol{\beta})\big[f_0(\boldsymbol{x}_j^T\boldsymbol{\beta}_0)-G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]\\&-G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})n^{-1}\sum_{j=1,j\neq i}^{n}K_h(\boldsymbol{x}_i^T\boldsymbol{\beta}-\boldsymbol{x}_j^T\boldsymbol{\beta})\\\triangleq&A_{n21}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-A_{n22}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}),\\A_{n3}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})=&(n-1)^{-1}\sum_{i=1}^{n}K_h'(\boldsymbol{x}_i^T\boldsymbol{\beta}-\boldsymbol{x}_j^T\boldsymbol{\beta}).\end{aligned}$$

First, similarly as in the proof of Lemma B4, we can derive that

$$P\left(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}}|A_{n1}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})|\geqslant c_0h\right)\leqslant\exp(-c_1nh^5).$$

We denote the p.d.f of $\boldsymbol{x}^T\boldsymbol{\beta}$ as $f_{\boldsymbol{\beta}}(\cdot)$. Note that $\boldsymbol{x}_j^T\boldsymbol{\beta}$ is independent of $\boldsymbol{x}_i^T\boldsymbol{\beta}$. We thus have

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{x}_i^T\boldsymbol{\beta}}\left[K_h'(\boldsymbol{x}_i^T\boldsymbol{\beta}-\boldsymbol{x}_j^T\boldsymbol{\beta})\,\middle|\boldsymbol{x}_i^T\boldsymbol{\beta}\right] &= h^{-2}\int K'\Big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}-y}{h}\Big)f_{\boldsymbol{\beta}}(y)dy\\
&= h^{-1}\int K'(-z)\Big[f_{\boldsymbol{\beta}}(\boldsymbol{x}_i^T\boldsymbol{\beta})+hzf_{\boldsymbol{\beta}}'(\boldsymbol{x}_i^T\boldsymbol{\beta})+\frac{h^2z^2}{2}f_{\boldsymbol{\beta}}''(\widetilde{t})\Big]dz\\
&= f_{\boldsymbol{\beta}}'(\boldsymbol{x}_i^T\boldsymbol{\beta})+\frac{h}{2}\int z^2K'(-z)f_{\boldsymbol{\beta}}''(\widetilde{t})dz,
\end{aligned}$$

where $\widetilde{t}$ is between $\boldsymbol{x}_i^T\boldsymbol{\beta}$ and $\boldsymbol{x}_i^T\boldsymbol{\beta}+hz$. In the above, the second equality considers Taylor expansion at point $\boldsymbol{x}_j^T\boldsymbol{\beta}$, with the notation $z=-(\boldsymbol{x}_j^T\boldsymbol{\beta}-y)/h$, which is followed by $y=\boldsymbol{x}_i^T\boldsymbol{\beta}+hz$. Similarly as in the proof of Lemma B6, we can show that for universal positive constants c_0 and c_1,

$$P\left(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}}\left|A_{n3}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-f_{\boldsymbol{\beta}}'(\boldsymbol{x}_i^T\boldsymbol{\beta})\right|\geqslant c_0h\right)\leqslant\exp(-c_1nh^5).$$

Then the techniques in the proof of Lemma B4 and Lemma B5 can be applied to analyze

$A_{n21}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})$ and $A_{n22}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})$. Observe that

$$
\begin{aligned}
&\mathrm{E}\Big\{K_h'(\boldsymbol{x}_i^T\boldsymbol{\beta}-\boldsymbol{x}_j^T\boldsymbol{\beta})\big[f_0(\boldsymbol{x}_j^T\boldsymbol{\beta}_0)-G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]\Big\}\\
=&\mathrm{E}_{(\boldsymbol{x}_i^T\boldsymbol{\beta},\boldsymbol{x}_j^T\boldsymbol{\beta})}\left\{\mathrm{E}\left\{K_h'(\boldsymbol{x}_i^T\boldsymbol{\beta}-\boldsymbol{x}_j^T\boldsymbol{\beta})\big[f_0(\boldsymbol{x}_j^T\boldsymbol{\beta}_0)-G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]\Big|\boldsymbol{x}_i^T\boldsymbol{\beta},\boldsymbol{x}_j^T\boldsymbol{\beta}\right\}\right\}\\
=&\mathrm{E}_{(\boldsymbol{x}_i^T\boldsymbol{\beta},\boldsymbol{x}_j^T\boldsymbol{\beta})}\Big\{K_h'(\boldsymbol{x}_i^T\boldsymbol{\beta}-\boldsymbol{x}_j^T\boldsymbol{\beta})\big[G(\boldsymbol{x}_j^T\boldsymbol{\beta}|\boldsymbol{\beta})-G(t|\boldsymbol{\beta})\big]\Big\}\\
=&h^{-1}\int K'(-z)\big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}+hz|\boldsymbol{\beta})-G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]f_{\boldsymbol{\beta}}(\boldsymbol{x}_i^T\boldsymbol{\beta}+hz)dz\\
=&-\int K'(z)\Big[zG^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})+\frac{hz^2}{2}G^{(2)}(t_1|\boldsymbol{\beta})\Big]\Big[f_{\boldsymbol{\beta}}(\boldsymbol{x}_i^T\boldsymbol{\beta})+hzf_{\boldsymbol{\beta}}'(t_2)\Big]dz\\
=&G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})f_{\boldsymbol{\beta}}(\boldsymbol{x}_i^T\boldsymbol{\beta})-\frac{hf_{\boldsymbol{\beta}}(\boldsymbol{x}_i^T\boldsymbol{\beta})}{2}\int z^2K'(z)G^{(2)}(t_1|\boldsymbol{\beta})dz\\
&-hG^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\int z^2K'(z)f_{\boldsymbol{\beta}}'(t_2)dz-\frac{h^2}{2}\int z^3K'(z)G^{(2)}(t_1|\boldsymbol{\beta})f_{\boldsymbol{\beta}}'(t_2)dz,
\end{aligned}
$$

where t_1 and t_2 are both between $\boldsymbol{x}_i^T\boldsymbol{\beta}$ and $\boldsymbol{x}_i^T\boldsymbol{\beta}+hz$. In the above, the second equality applies the independence between $\boldsymbol{x}_i^T\boldsymbol{\beta}$ and $\boldsymbol{x}_j^T\boldsymbol{\beta}$, and $G(\boldsymbol{x}_j^T\boldsymbol{\beta}|\boldsymbol{\beta})=\mathrm{E}\big[f_0(\boldsymbol{x}_j^T\boldsymbol{\beta}_0)|\boldsymbol{x}_j^T\boldsymbol{\beta}\big]$. Then Assumption (A3)–(A5) and the proofs in Lemma B5 and Lemma B6 imply that for some constants c_0, c_1, and all n sufficiently large,

$$
P\left(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}}\big|A_{n21}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})f_{\boldsymbol{\beta}}(\boldsymbol{x}_i^T\boldsymbol{\beta})\big|\geqslant c_0h\right)\leqslant\exp[-c_1\log(p\vee n)],
$$

$$
P\left(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}}\big|A_{n22}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})f_{\boldsymbol{\beta}}(\boldsymbol{x}_i^T\boldsymbol{\beta})\big|\geqslant c_0h\right)\leqslant\exp[-c_1\log(p\vee n)].
$$

This implies that $P\Big(\max_{1\leqslant i\leqslant n}\sup\limits_{\boldsymbol{\beta}\in\mathbb{B}}\big|A_{n2}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big|\leqslant c_0h\Big)\geqslant 1-2\exp[-c_1\log(p\vee n)]$. Assumption (A4) implies that $\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}}f_{\boldsymbol{\beta}}^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta})\leqslant M$, for some positive constant M. by noting the high probability bounds for $A_{ni}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})$ and $A_{ni}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})$, we conclude the lemma. □

Proof of Lemma A7. We will prove the first part of the claim below. The proof of the second and third parts is similar.

Note that

$$\widehat{\mathrm{E}}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}) - \mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}) = \frac{(n-1)^{-1}\sum_{j=1,j\neq i}^{n} K_h(\boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{x}_j^T\boldsymbol{\beta})\big[\boldsymbol{x}_j - \mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})\big]}{A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})}.$$

Let $B_n(\boldsymbol{x}_i^T\boldsymbol{\beta}, \boldsymbol{v}|\boldsymbol{\beta}) = [(n-1)h]^{-1}\sum_{j=1,j\neq i}^{n} \boldsymbol{\gamma}_i(z_j)^T\boldsymbol{v}$, where $\boldsymbol{\gamma}_i(z_j) = K\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}-\boldsymbol{x}_j^T\boldsymbol{\beta}}{h}\big)\big[\boldsymbol{x}_j - \mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})\big]$. Lemma B6 and inequality (S26) already provide a high probability bound for the denominator. It suffices to prove the high probability bound for $\sup_{\substack{\boldsymbol{\beta}\in\mathbb{B} \\ \boldsymbol{v}\in\mathbb{K}(p,2ks)}} |B_n(\boldsymbol{x}_i^T\boldsymbol{\beta}, \boldsymbol{v}|\boldsymbol{\beta})|$.

We first derive the bound of $\big|\mathrm{E}\big[B_n(\boldsymbol{x}_i^T\boldsymbol{\beta}, \boldsymbol{v}|\boldsymbol{\beta})\big]\big|$. Let $f_{\boldsymbol{\beta}}(\cdot)$ denote the p.d.f of $\boldsymbol{x}^T\boldsymbol{\beta}$. Note

$$\begin{aligned}
\mathrm{E}[\boldsymbol{\gamma}_i(z_j)^T\boldsymbol{v}] =& \mathrm{E}\Big\{K\Big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{x}_j^T\boldsymbol{\beta}}{h}\Big)\big[\boldsymbol{x}_j - \mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})\big]^T\boldsymbol{v}\Big\} \\
=& \mathrm{E}_{(\boldsymbol{x}_i^T\boldsymbol{\beta}, \boldsymbol{x}_j^T\boldsymbol{\beta})}\Big\{\mathrm{E}\Big[K\Big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{x}_j^T\boldsymbol{\beta}}{h}\Big)\big[\boldsymbol{x}_j - \mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})\big]^T\boldsymbol{v}|\boldsymbol{x}_i^T\boldsymbol{\beta}, \boldsymbol{x}_j^T\boldsymbol{\beta}\Big]\Big\} \\
=& \mathrm{E}_{(\boldsymbol{x}_i^T\boldsymbol{\beta}, \boldsymbol{x}_j^T\boldsymbol{\beta})}\Big\{K\Big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{x}_j^T\boldsymbol{\beta}}{h}\Big)\big[\mathrm{E}(\boldsymbol{x}_j|\boldsymbol{x}_j^T\boldsymbol{\beta}) - \mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})\big]^T\boldsymbol{v}\Big\} \\
=& \mathrm{E}_{\boldsymbol{x}_i^T\boldsymbol{\beta}}\left\{h\int K(-z)\big[\mathrm{E}(\boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_i^T\boldsymbol{\beta} + hz) - \mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})\big]^T\boldsymbol{v}f_{\boldsymbol{\beta}}(\boldsymbol{x}_i^T\boldsymbol{\beta} + hz)dz\right\} \\
=& \mathrm{E}_{\boldsymbol{x}_i^T\boldsymbol{\beta}}\Big\{h\int K(-z)\Big[\mathrm{E}^{(1)}(\boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_i^T\boldsymbol{\beta})hz \\
&+ \frac{h^2z^2}{2}\mathrm{E}^{(2)}(\boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\beta} = t_1)\Big]^T\boldsymbol{v}\Big[f_{\boldsymbol{\beta}}(\boldsymbol{x}_i^T\boldsymbol{\beta}) + hzf'_{\boldsymbol{\beta}}(\widetilde{t})\Big]\Big\}dz \\
=& \mathrm{E}_{\boldsymbol{x}_i^T\boldsymbol{\beta}}\left\{\frac{h^3f_{\boldsymbol{\beta}}(\boldsymbol{x}_i^T\boldsymbol{\beta})}{2}\int z^2K(-z)\mathrm{E}^{(2)}(\boldsymbol{x}^T\boldsymbol{v}|\boldsymbol{x}^T\boldsymbol{\beta} = t_1)dz\right\} \\
&+ \mathrm{E}_{\boldsymbol{x}_i^T\boldsymbol{\beta}}\left\{h^3\mathrm{E}^{(1)}(\boldsymbol{x}^T\boldsymbol{v}|\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_i^T\boldsymbol{\beta})\int z^2K(-z)f'_{\boldsymbol{\beta}}(\widetilde{t})dz\right\} \\
&+ \mathrm{E}_{\boldsymbol{x}_i^T\boldsymbol{\beta}}\left\{\frac{h^4}{2}\int z^3K(-z)\mathrm{E}^{(2)}(\boldsymbol{x}^T\boldsymbol{v}|\boldsymbol{x}^T\boldsymbol{\beta} = t_1)f'_{\boldsymbol{\beta}}(\widetilde{t})dz\right\},
\end{aligned}$$

where t_1 and $\widetilde{t}$ are both between $\boldsymbol{x}_i^T\boldsymbol{\beta}$ and $\boldsymbol{x}_i^T\boldsymbol{\beta} + hz$. In the above, the third equality uses the independence between $\boldsymbol{x}_i^T\boldsymbol{\beta}$ and $\boldsymbol{x}_j^T\boldsymbol{\beta}$. The last equality uses $\int zK(-z)dz = 0$. Assumptions (A2)–(A4) imply that $\big|\mathrm{E}\big[B_n(\boldsymbol{x}_i^T\boldsymbol{\beta}, \boldsymbol{v}|\boldsymbol{\beta})\big]\big| \leqslant ch^2||\boldsymbol{v}||_2$ for some positive constant c. Applying the same techniques as those in the proof of Lemma B4, we can derive that $\mathrm{E}\{[\boldsymbol{\gamma}_i(z_j)^T\boldsymbol{v}]^2\} \leqslant ch||\boldsymbol{v}||_2^2$.

Next, we derive the high probability bound of $|B_n(\boldsymbol{x}_i^T\boldsymbol{\beta}, \boldsymbol{v}|\boldsymbol{\beta})|$. Note that $\boldsymbol{x}_j^T\boldsymbol{v}$ is independent of $\boldsymbol{x}_i^T\boldsymbol{\beta}$. The sub-Gaussian property of $\boldsymbol{x}_j^T\boldsymbol{v}$ and $\mathrm{E}(\boldsymbol{x}_i^T\boldsymbol{v}|\boldsymbol{x}_i^T\boldsymbol{\beta})$ implies that $\boldsymbol{x}_j^T\boldsymbol{v} - \mathrm{E}(\boldsymbol{x}_i^T\boldsymbol{v}|\boldsymbol{x}_i^T\boldsymbol{\beta})$ is also sub-Gaussian. Since $K(\cdot)$ is bounded on the real line, Lemma A2 implies that $[\boldsymbol{\gamma}_i(z_j) - \mathrm{E}\boldsymbol{\gamma}_i(z_j)]^T\boldsymbol{v}$ is sub-Gaussian. Then the tail probability inequality for sub-Gaussian implies that

$$P\left(\left|\sum_{j=1, j\neq i}^{n} [\boldsymbol{\gamma}_i(z_j) - \mathrm{E}\boldsymbol{\gamma}_i(z_j)]^T\boldsymbol{v}\right| \geqslant t \;\middle|\; \boldsymbol{x}_i^T\boldsymbol{\beta}\right) \leqslant 2\exp\left[-\frac{t^2}{2c^2(n-1)||\boldsymbol{v}||_2^2 h}\right],$$

for some positive constant c, where applies $\mathrm{E}\{[\boldsymbol{\gamma}_i(z_j)^T\boldsymbol{v}]^2\} \leqslant c^2 h||\boldsymbol{v}||_2^2$. Taking $t = c(n-1)h^3$, we have that

$$\begin{aligned}
&P\left(\left|\sum_{j=1, j\neq i}^{n} [\boldsymbol{\gamma}_i(z_j) - \mathrm{E}\boldsymbol{\gamma}_i(z_j)]^T\boldsymbol{v}\right| \geqslant c(n-1)h^3\right)\\
=&\mathrm{E}_{\boldsymbol{x}_i^T\boldsymbol{\beta}}\left\{P\left(\left|\sum_{j=1, j\neq i}^{n} [\boldsymbol{\gamma}_i(z_j) - \mathrm{E}\boldsymbol{\gamma}_i(z_j)]^T\boldsymbol{v}\right| \geqslant c(n-1)h^3 \;\middle|\; \boldsymbol{x}_i^T\boldsymbol{\beta}\right)\right\}\\
\leqslant& 2\exp(-cnh^5),
\end{aligned}$$

for some positive constant c and all n sufficiently large, since $||\boldsymbol{v}||_2 \leqslant 1$. Combining this with the bound of $\left|\mathrm{E}\big[B_n(\boldsymbol{x}_i^T\boldsymbol{\beta}, \boldsymbol{v}|\boldsymbol{\beta})\big]\right|$, we conclude that there exist some positive constants c_0 and c_1 such that for all n sufficiently large,

$$P\Big(\big|B_n(\boldsymbol{x}_i^T\boldsymbol{\beta}, \boldsymbol{v}|\boldsymbol{\beta})\big| \geqslant c_0 h^2\Big) \leqslant \exp(-c_1 nh^5).$$

To obtain the uniform bound, we will cover $\mathbb{B}$ with N_1 L_2-balls of radius δ. Denote the centers by $\boldsymbol{\beta}_1^*, \cdots, \boldsymbol{\beta}_{N_1}^*$. Similarly in the proof of Lemma B4, we can cover $\mathbb{K}(p, 2ks)$ with N_2 L_2-balls of radius δ. Denote their centers by $\boldsymbol{v}_1^*, \cdots, \boldsymbol{v}_{N_2}^*$. Let $\mathcal{N}_\delta$ be the this joint cover of $\mathbb{B} \times \mathbb{K}(p, 2ks)$. We can construct the covers such that $N \triangleq |\mathcal{N}_\delta| = N_1 * N_2 \leqslant cp^{4ks}\delta^{-4ks}$ for some positive constant c. Given any $\boldsymbol{\beta} \in \mathbb{B}$ and $\boldsymbol{v} \in \mathbb{K}(p, 2ks)$, we can find $(\boldsymbol{\beta}^*, \boldsymbol{v}^*) \in \mathcal{N}_\delta$

such that $||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2 \leqslant \delta$, and $||\boldsymbol{v}-\boldsymbol{v}^*||_2 \leqslant \delta$. We have

$$
\begin{aligned}
&\Big|(n-1)^{-1}\sum_{j=1,j\neq i}^{n} K_h(\boldsymbol{x}_i^T\boldsymbol{\beta}-\boldsymbol{x}_j^T\boldsymbol{\beta})\big[\boldsymbol{x}_j-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})\big]^T\boldsymbol{v}\\
&\quad -(n-1)^{-1}\sum_{j=1,j\neq i}^{n} K_h(\boldsymbol{x}_i^T\boldsymbol{\beta}^*-\boldsymbol{x}_j^T\boldsymbol{\beta}^*)\big[\boldsymbol{x}_j-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}^*)\big]^T\boldsymbol{v}^*\Big|\\
\leqslant&\Big|(n-1)^{-1}\sum_{j=1,j\neq i}^{n}\Big[K_h(\boldsymbol{x}_i^T\boldsymbol{\beta}-\boldsymbol{x}_j^T\boldsymbol{\beta})-K_h(\boldsymbol{x}_i^T\boldsymbol{\beta}^*-\boldsymbol{x}_j^T\boldsymbol{\beta}^*)\Big]\big[\boldsymbol{x}_j-\mathrm{E}(\boldsymbol{x}_j|\boldsymbol{x}_j^T\boldsymbol{\beta})\big]^T\boldsymbol{v}\Big|\\
&\quad+\Big|(n-1)^{-1}\sum_{j=1,j\neq i}^{n}\Big[K_h(\boldsymbol{x}_i^T\boldsymbol{\beta}-\boldsymbol{x}_j^T\boldsymbol{\beta})-K_h(\boldsymbol{x}_i^T\boldsymbol{\beta}^*-\boldsymbol{x}_j^T\boldsymbol{\beta}^*)\Big]\big[\mathrm{E}(\boldsymbol{x}_j|\boldsymbol{x}_j^T\boldsymbol{\beta})-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})\big]^T\boldsymbol{v}\Big|\\
&\quad+\Big|(n-1)^{-1}\sum_{j=1,j\neq i}^{n} K_h(\boldsymbol{x}_i^T\boldsymbol{\beta}^*-\boldsymbol{x}_j^T\boldsymbol{\beta}^*)\big[\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}^*)-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})\big]^T\boldsymbol{v}\Big|\\
&\quad+\Big|(n-1)^{-1}\sum_{j=1,j\neq i}^{n} K_h(\boldsymbol{x}_i^T\boldsymbol{\beta}^*-\boldsymbol{x}_j^T\boldsymbol{\beta}^*)\big[\boldsymbol{x}_j-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}^*)\big]^T(\boldsymbol{v}-\boldsymbol{v}^*)\Big|\\
\triangleq&\sum_{i=1}^{4}|I_{ni}|,
\end{aligned}
$$

where the definition of I_{ni} is clear from the context. Lemma B2 implies that

$$P\left(\max_{1\leqslant i\leqslant n}|\boldsymbol{x}_i^T\boldsymbol{\beta}|\geqslant\sigma_x\sqrt{s\log(p\vee n)}||\boldsymbol{\beta}||_2,\ \forall\ \boldsymbol{\beta}\in\mathbb{B}\right)\leqslant\exp[-cs\log(p\vee n)], \tag{S27}$$

$$P\left(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{v}\in\mathbb{K}(p,4ks)}\big|\mathrm{E}(\boldsymbol{x}_i^T\boldsymbol{v}|\boldsymbol{x}_i^T\boldsymbol{\beta})\big|\geqslant\sigma_x\sqrt{s\log(p\vee n)},\ \forall\ \boldsymbol{\beta}\in\mathbb{B}\right)\leqslant\exp[-cs\log(p\vee n)], \tag{S28}$$

$$P\left(\max_{1\leqslant i\leqslant n}|\boldsymbol{x}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}^*)|\geqslant\sigma_x\sqrt{s\log(p\vee n)}||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2,\ \forall\ \boldsymbol{\beta},\boldsymbol{\beta}^*\in\mathbb{B}\right)\leqslant\exp[-cs\log(p\vee n)], \tag{S29}$$

for some positive constant c, and all n sufficiently large, where the analysis of (S28) is similar as the proofs of Lemma A3 and Lemma B7. Since $\boldsymbol{x}_j^T(\boldsymbol{\beta}-\boldsymbol{\beta}^*)$ are independent sub-Gaussian random variables, Lemma B2 implies that

$$P\Bigg((n-1)^{-1}\left|\sum_{j=1,j\neq i}^{n}|\boldsymbol{x}_j^T(\boldsymbol{\beta}-\boldsymbol{\beta}^*)|^2-(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^T\mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T)(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\right|$$
$$\geqslant c_0\sigma_x^2\sqrt{\frac{s\log(p\vee n)}{n}}||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2^2, \forall\ \boldsymbol{\beta},\boldsymbol{\beta}^*\in\mathbb{B}\Bigg)$$
$$\leqslant \exp[-cs\log(p\vee n)],$$

for some positive constants c_0, c, and all n sufficiently large. Assumption (A2) implies that $(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^T\mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T)(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\leqslant \xi_3||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2^2$. Hence we have that

$$\begin{aligned}(n-1)^{-1}\sum_{j=1,j\neq i}^{n}|\boldsymbol{x}_j^T(\boldsymbol{\beta}-\boldsymbol{\beta}^*)|^2 &\leqslant \left(\xi_3+c_0\sigma_x^2\sqrt{\frac{s\log(p\vee n)}{n}}\right)||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2^2\\ &\leqslant \sigma_x^2 s\log(p\vee n)||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2^2,\ \forall\ \boldsymbol{\beta},\boldsymbol{\beta}^*\in\mathbb{B}\end{aligned}\tag{S30}$$

with probability at least $1-\exp[-c_1\log(p\vee n)]$, for some positive constants c_0, c_1, and all n sufficiently large. Similarly the sub-Gaussian property of $\mathrm{E}(\boldsymbol{x}_j^T\boldsymbol{v}|\boldsymbol{x}_j^T\boldsymbol{\beta})$ and $\boldsymbol{x}_j^T\boldsymbol{v}-\mathrm{E}(\boldsymbol{x}_j^T\boldsymbol{v}|\boldsymbol{x}_j^T\boldsymbol{\beta})$, Lemma B1 and Lemma B2 imply that

$$P\left(\sup_{\boldsymbol{v}\in\mathbb{K}(p,4ks)}\frac{1}{n-1}\sum_{j=1,j\neq i}^{n}\left[\mathrm{E}(\boldsymbol{x}_j^T\boldsymbol{v}|\boldsymbol{x}_j^T\boldsymbol{\beta})\right]^2\geqslant c_0\sigma_x^2,\ \forall\ \boldsymbol{\beta}\in\mathbb{B}\right)\leqslant\exp[-c\log(p\vee n)],\tag{S31}$$

$$P\left(\sup_{\boldsymbol{v}\in\mathbb{K}(p,4ks)}\frac{1}{n-1}\sum_{j=1,j\neq i}^{n}\left(\boldsymbol{x}_j^T\boldsymbol{v}\right)^2\geqslant c_0\sigma_x^2\right)\leqslant\exp[-c\log(p\vee n)],\tag{S32}$$

$$P\left(\sup_{\boldsymbol{v}\in\mathbb{K}(p,4ks)}\frac{1}{n-1}\sum_{j=1,j\neq i}^{n}\left[\boldsymbol{x}_j^T\boldsymbol{v}-\mathrm{E}(\boldsymbol{x}_j^T\boldsymbol{v}|\boldsymbol{x}_j^T\boldsymbol{\beta})\right]^2\geqslant c_0\sigma_x^2,\ \forall\ \boldsymbol{\beta}\in\mathbb{B}\right)\leqslant\exp[-c\log(p\vee n)],\tag{S33}$$

for some positive constants c_0, c, and all n sufficiently large, where the analysis of (S31) and

(S33) is similar as the proofs of Lemma A3 and Lemma B7. Denote the event

$$\mathcal{E} = \Big\{ \max_{1\leqslant i\leqslant n} |\boldsymbol{x}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}^*)|^2 + \frac{1}{n-1}\sum_{j=1,j\neq i}^{n} |\boldsymbol{x}_j^T(\boldsymbol{\beta}-\boldsymbol{\beta}^*)|^2 \leqslant 2\sigma_x^2 s\log(p\vee n)||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2^2,\ \forall\ \boldsymbol{\beta},\boldsymbol{\beta}^*\in\mathbb{B}\Big\}$$

$$\bigcap\Big\{ \max_{1\leqslant i\leqslant n} \sup_{\substack{\boldsymbol{\beta}\in\mathbb{B}\\ \boldsymbol{v}\in\mathbb{K}(p,4ks)}} \Big[\big|\mathrm{E}(\boldsymbol{x}_i^T\boldsymbol{v}|\boldsymbol{x}_i^T\boldsymbol{\beta})\big|^2 + \frac{1}{n-1}\sum_{j=1,j\neq i}^{n}\big|\mathrm{E}(\boldsymbol{x}_j^T\boldsymbol{v}|\boldsymbol{x}_j^T\boldsymbol{\beta})\big|^2\Big] \leqslant 2c_0\sigma_x^2 s\log(p\vee n)\Big\}$$

$$\bigcap\Big\{ \sup_{\substack{\boldsymbol{\beta}\in\mathbb{B}\\ \boldsymbol{v}\in\mathbb{K}(p,4ks)}} \frac{1}{n-1}\sum_{j=1,j\neq i}^{n}\big[\boldsymbol{x}_j^T\boldsymbol{v}-\mathrm{E}(\boldsymbol{x}_j^T\boldsymbol{v}|\boldsymbol{x}_j^T\boldsymbol{\beta})\big]^2 \leqslant 2\sigma_x^2\Big\}$$

$$\bigcap\Big\{ \sup_{\boldsymbol{v}\in\mathbb{K}(p,4ks)} \frac{1}{n-1}\sum_{j=1,j\neq i}^{n}(\boldsymbol{x}_j^T\boldsymbol{v})^2 \leqslant 2\sigma_x^2\Big\}\bigcap\Big\{\max_{1\leqslant i\leqslant n}|\boldsymbol{x}_i^T\boldsymbol{\beta}| \leqslant \sigma_x\sqrt{s\log(p\vee n)}||\boldsymbol{\beta}||_2,\ \forall\ \boldsymbol{\beta}\in\mathbb{B}\Big\}.$$

Combining (S27) – (S33), we have $P(\mathcal{E}) \geqslant 1 - 7\exp[-c_1 s\log(p\vee n)]$, for some positive constants c_0, c_1, and all n sufficiently large.

Take $\delta = \frac{h^4}{8s\log(p\vee n)}$. According to Assumption (A3), we observe

$$\begin{aligned}
|I_{n1}| \leqslant & \sqrt{(n-1)^{-1}\sum_{j=1,j\neq i}^{n}\Big|K_h(\boldsymbol{x}_i^T\boldsymbol{\beta}-\boldsymbol{x}_j^T\boldsymbol{\beta}) - K_h(\boldsymbol{x}_i^T\boldsymbol{\beta}^*-\boldsymbol{x}_j^T\boldsymbol{\beta}^*)\Big|^2} \\
& * \sqrt{(n-1)^{-1}\sum_{j=1,j\neq i}^{n}\Big|\big[\boldsymbol{x}_j-\mathrm{E}(\boldsymbol{x}_j|\boldsymbol{x}_j^T\boldsymbol{\beta})\big]^T\boldsymbol{v}\Big|^2} \\
\leqslant & 2h^{-2}\sqrt{|\boldsymbol{x}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}^*)|^2 + (n-1)^{-1}\sum_{j=1,j\neq i}^{n}|\boldsymbol{x}_j^T(\boldsymbol{\beta}-\boldsymbol{\beta}^*)|^2} \\
& * \sqrt{(n-1)^{-1}\sum_{j=1,j\neq i}^{n}\big[\boldsymbol{x}_j^T\boldsymbol{v}-\mathrm{E}(\boldsymbol{x}_j^T\boldsymbol{v}|\boldsymbol{x}_j^T\boldsymbol{\beta})\big]^2} \\
\leqslant & c_0 h^{-2}\delta||\boldsymbol{v}||_2\sqrt{s\log(p\vee n)} \leqslant \frac{c_0 h^2}{8\sqrt{s\log(p\vee n)}},
\end{aligned}$$

on the event $\mathcal{E}$, for some constant $c_0 > 0$, and all n sufficiently large, since $||\boldsymbol{v}||_2 \leqslant 1$. In the above, the last inequality applies the first and the third events in $\mathcal{E}$.

Similarly for I_{n2}, we have

$$
\begin{aligned}
|I_{n2}| \leqslant & \sqrt{(n-1)^{-1} \sum_{j=1, j \neq i}^{n} \Big| K_h(\boldsymbol{x}_i^T \boldsymbol{\beta} - \boldsymbol{x}_j^T \boldsymbol{\beta}) - K_h(\boldsymbol{x}_i^T \boldsymbol{\beta}^* - \boldsymbol{x}_j^T \boldsymbol{\beta}^*) \Big|^2} \\
& * \sqrt{(n-1)^{-1} \sum_{j=1, j \neq i}^{n} \Big| \big[\mathrm{E}(\boldsymbol{x}_j | \boldsymbol{x}_j^T \boldsymbol{\beta}) - \mathrm{E}(\boldsymbol{x}_i | \boldsymbol{x}_i^T \boldsymbol{\beta}) \big]^T \boldsymbol{v} \Big|^2} \\
\leqslant & 2h^{-2} \sqrt{|\boldsymbol{x}_i^T (\boldsymbol{\beta} - \boldsymbol{\beta}^*)|^2 + (n-1)^{-1} \sum_{j=1, j \neq i}^{n} |\boldsymbol{x}_j^T (\boldsymbol{\beta} - \boldsymbol{\beta}^*)|^2} \\
& * \sqrt{\big| \mathrm{E}(\boldsymbol{x}_i^T \boldsymbol{v} | \boldsymbol{x}_i^T \boldsymbol{\beta}) \big|^2 + (n-1)^{-1} \sum_{j=1, j \neq i}^{n} \big| \mathrm{E}(\boldsymbol{x}_j^T \boldsymbol{v} | \boldsymbol{x}_j^T \boldsymbol{\beta}) \big|^2} \\
\leqslant & c_0 h^{-2} \delta ||\boldsymbol{v}||_2 s \log(p \vee n) \leqslant c_0 h^2 / 8,
\end{aligned}
$$

on the event $\mathcal{E}$, for some constant $c_0 > 0$, and all n sufficiently large. In the above, the last inequality applies the first and the second events in $\mathcal{E}$.

Assumption (A2)-(c) implies

$$
\begin{aligned}
|I_{n3}| & \leqslant ch^{-1} \Big| \big[\mathrm{E}(\boldsymbol{x}_i | \boldsymbol{x}_i^T \boldsymbol{\beta}^*) - \mathrm{E}(\boldsymbol{x}_i | \boldsymbol{x}_i^T \boldsymbol{\beta}) \big]^T \boldsymbol{v} \Big| \\
& \leqslant ch^{-1} ||\boldsymbol{v}||_2 \Big[|\boldsymbol{x}_i^T (\boldsymbol{\beta} - \boldsymbol{\beta}^*)| + (|\boldsymbol{x}_i^T \boldsymbol{\beta}| + |\boldsymbol{x}_i^T \boldsymbol{\beta}^*|) * ||\boldsymbol{\beta} - \boldsymbol{\beta}^*||_2 \Big] \\
& \leqslant c_0 h^{-1} \delta ||\boldsymbol{v}||_2 \sqrt{s \log(p \vee n)} \leqslant \frac{c_0 h^3}{8 \sqrt{s \log(p \vee n)}},
\end{aligned}
$$

on the event $\mathcal{E}$, for some positive constants c, c_0, and all n sufficiently large. In the above, the second last inequality applies the first and the last events in $\mathcal{E}$.

Assumption (A3) implies that

$$
\begin{aligned}
|I_{n4}| &\leqslant ch^{-1}(n-1)^{-1}\sum_{j=1,j\neq i}^{n}\Big|\big[\boldsymbol{x}_j-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}^*)\big]^T(\boldsymbol{v}-\boldsymbol{v}^*)\Big| \\
&\leqslant \sqrt{2}ch^{-1}\sqrt{\big[\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}^*)^T(\boldsymbol{v}-\boldsymbol{v}^*)\big]^2+(n-1)^{-1}\sum_{j=1,j\neq i}^{n}\big[\boldsymbol{x}_j^T(\boldsymbol{v}-\boldsymbol{v}^*)\big]^2} \\
&\leqslant c_0h^{-1}\sqrt{s\log(p\vee n)}||\boldsymbol{v}-\boldsymbol{v}^*||_2\leqslant\frac{c_0h^3}{8\sqrt{s\log(p\vee n)}},
\end{aligned}
$$

on the event $\mathcal{E}$, for some positive constants c, c_0, and all n sufficiently large. In the above, the second last inequality applies the second and the fourth events in $\mathcal{E}$, since $\big[\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}^*)^T(\boldsymbol{v}-\boldsymbol{v}^*)\big]^2 \leqslant \big|\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})^T(\boldsymbol{v}-\boldsymbol{v}^*)\big|^2+\frac{1}{n-1}\sum_{j=1,j\neq i}^{n}\big|\mathrm{E}(\boldsymbol{x}_j|\boldsymbol{x}_j^T\boldsymbol{\beta})^T(\boldsymbol{v}-\boldsymbol{v}^*)\big|^2$, for any $\boldsymbol{v}-\boldsymbol{v}^*\in\mathbb{K}(p,4ks)$.

Combining the above results, we conclude that on the event $\mathcal{E}$,

$$
\begin{aligned}
&\Big|(n-1)^{-1}\sum_{j=1,j\neq i}^{n}K_h(\boldsymbol{x}_i^T\boldsymbol{\beta}-\boldsymbol{x}_j^T\boldsymbol{\beta})\big[\boldsymbol{x}_j-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})\big]^T\boldsymbol{v} \\
&-(n-1)^{-1}\sum_{j=1,j\neq i}^{n}K_h(\boldsymbol{x}_i^T\boldsymbol{\beta}^*-\boldsymbol{x}_j^T\boldsymbol{\beta}^*)\big[\boldsymbol{x}_j-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}^*)\big]^T\boldsymbol{v}^*\Big| \\
\leqslant&\frac{c_0h^2}{2},
\end{aligned}
$$

for some positive constant c_0, and all n sufficiently large. Hence,

$$
\begin{aligned}
&P\Big(\sup_{\substack{\boldsymbol{\beta}\in\mathbb{B}\\ \boldsymbol{v}\in\mathbb{K}(p,2ks)}} |B_n(\boldsymbol{x}_i^T\boldsymbol{\beta},\boldsymbol{v}|\boldsymbol{\beta})| \geqslant c_0h^2\Big)\\
\leqslant &P\Big(\bigcup_{(\boldsymbol{\beta}^*,\boldsymbol{v}^*)\in\mathcal{N}_\delta} \big|B_n(\boldsymbol{x}_i^T\boldsymbol{\beta}^*,\boldsymbol{v}^*|\boldsymbol{\beta}^*)\big| \geqslant c_0h^2/2\Big)\\
&+P\Big(\sup_{(\boldsymbol{\beta}^*,\boldsymbol{v}^*)\in\mathcal{N}_\delta}\sup_{\substack{||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2\leqslant\delta\\ ||\boldsymbol{v}-\boldsymbol{v}^*||_2\leqslant\delta}} |B_n(\boldsymbol{x}_i^T\boldsymbol{\beta},\boldsymbol{v}|\boldsymbol{\beta})-B_n(\boldsymbol{x}_i^T\boldsymbol{\beta}^*,\boldsymbol{v}^*|\boldsymbol{\beta}^*)| \geqslant c_0h^2/2\Big)\\
\leqslant &\sum_{(\boldsymbol{\beta}^*,\boldsymbol{v}^*)\in\mathcal{N}_\delta} P\Big(\big|B_n(\boldsymbol{x}_i^T\boldsymbol{\beta}^*,\boldsymbol{v}^*|\boldsymbol{\beta}^*)\big| \geqslant c_0h^2/2\Big)+7\exp[-c_1 s\log(p\vee n)]\\
\leqslant &cp^{4ks}\delta^{-4ks}\exp(-c_1nh^5)+7\exp[-c_1s\log(p\vee n)]\\
= &\exp[-c_2\log(p\vee n)],
\end{aligned}
$$

for some positive constants c_0, c_1, $c_2>1$, and all n sufficiently large,

Hence there exist some universal positive constants d_0 and d_1, such that for all n sufficiently large,

$$
\begin{aligned}
P\Big(\max_{1\leqslant i\leqslant n}\sup_{\substack{\boldsymbol{\beta}\in\mathbb{B}\\ \boldsymbol{v}\in\mathbb{K}(p,2ks)}} |B_n(\boldsymbol{x}_i^T\boldsymbol{\beta},\boldsymbol{v}|\boldsymbol{\beta})| \geqslant d_0h^2\Big) &\leqslant \sum_{i=1}^n P\Big(\sup_{\substack{\boldsymbol{\beta}\in\mathbb{B}\\ \boldsymbol{v}\in\mathbb{K}(p,2ks)}} |B_n(\boldsymbol{x}_i^T\boldsymbol{\beta},\boldsymbol{v}|\boldsymbol{\beta})| \geqslant d_0h^2\Big)\\
&\leqslant \exp[-d_1\log(p\vee n)],
\end{aligned}
$$

which concludes the lemma. $\square$

Proof of Lemma A8. We will prove (S7) below. The proof of (S8) is similar.

Theorem 1 implies that $P\left(\widehat{\boldsymbol{\beta}}\in\mathbb{B}_1\right)\geqslant 1-\exp(-c\log p)$, for some positive constant c, and all n sufficiently large. Lemma A11 implies that

$$
P\left(\max_{1\leqslant i\leqslant n}\left|\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})-G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right|\geqslant c_0h\right)\leqslant\exp(-c_1\log p),
$$

for some positive constants c_0, c_1, and all n sufficiently large. Define

$$\mathbb{M} =\Big\{m(\cdot|\boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathbb{B}_1,\ m(\cdot|\boldsymbol{\beta}) \in C_1^1(T),\ \forall\ \boldsymbol{\beta},\ \text{and}$$
$$\max_{1\leqslant i\leqslant n} \sup_{\boldsymbol{\beta}\in\mathbb{B}_1} \left|m(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right| \leqslant c_1 h\Big\},$$

for some positive constant c_1, where $T = \{t \in \mathbb{R} : |t| \leqslant 2||\boldsymbol{\beta}_0||_2\sigma_x\sqrt{\log(p \vee n)}\}$, and $C_1^1(T)$ is the set of all continuous and Lipschitz functions $f : T \to \mathbb{R}$. To prove (S7), it is sufficient to prove that there exist some positive constants c_0, c_1, such that for all n sufficiently large,

$$P\Big(\max_{2\leqslant j\leqslant p} \sup_{\boldsymbol{\beta}\in\mathbb{B}_1, m\in\mathbb{M}} \left|n^{-1/2}\sum_{i=1}^n \boldsymbol{\theta}_j^T\boldsymbol{\gamma}(Z_i, \boldsymbol{\beta}, m)\right| \geqslant c_0 \left[h^2 \log(p \vee n)\right]^{1/4}\Big) \leqslant \exp[-c_1 \log(p \vee n)],$$

where $Z_i = (\boldsymbol{x}_i, \epsilon_i, A_i)$, $\boldsymbol{\gamma}(Z_i, \boldsymbol{\beta}, m) = \left[m(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right]\widetilde{\epsilon}_i\left[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\right]$, $\widetilde{\epsilon}_i = 2(2A_i - 1)[\epsilon_i + g(\boldsymbol{x}_i)]$, and $m(\boldsymbol{x}^T\boldsymbol{\beta}|\boldsymbol{\beta})$ depends on $\boldsymbol{x}$ only through $\boldsymbol{x}^T\boldsymbol{\beta}$.

We have

$$\begin{aligned}
&P\left(\max_{2\leqslant j\leqslant p} \sup_{\boldsymbol{\beta}\in\mathbb{B}_1, m\in\mathbb{M}} \left|n^{-1/2}\sum_{i=1}^n \boldsymbol{\theta}_j^T\boldsymbol{\gamma}(Z_i, \boldsymbol{\beta}, m)\right| > t\right)\\
\leqslant &P\left(\max_{2\leqslant j\leqslant p} \sup_{\boldsymbol{\beta}\in\mathbb{B}_1, m\in\mathbb{M}} \left|n^{-1/2}\sum_{i=1}^n \boldsymbol{\theta}_j^T\boldsymbol{\gamma}(Z_i, \boldsymbol{\beta}, m)\right| > t \,\middle|\, \mathcal{H}_n \cap \mathcal{J}_n \cap \mathcal{K}_n\right) + \exp[-c \log(p \vee n)]\\
\leqslant &\sum_{j=2}^p P\left(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1, m\in\mathbb{M}} \left|n^{-1/2}\sum_{i=1}^n \boldsymbol{\theta}_j^T\boldsymbol{\gamma}(Z_i, \boldsymbol{\beta}, m)\right| > t \,\middle|\, \mathcal{H}_n \cap \mathcal{J}_n \cap \mathcal{K}_n\right) + \exp[-c \log(p \vee n)],
\end{aligned}$$

for some positive constant c and all n sufficiently large, where the events $\mathcal{H}_n$, $\mathcal{J}_n$, and $\mathcal{K}_n$ are defined in Lemma A3. In the proof below, we write $\boldsymbol{\gamma}(Z_i, \boldsymbol{\beta}, m)$ as $\boldsymbol{\gamma}_i(\boldsymbol{\beta}, m)$ for simplicity.

Note that $\boldsymbol{\theta}_j^T\boldsymbol{\gamma}_i(\boldsymbol{\beta}, m) = 2(2A_i - 1)\left[\epsilon_i + g(\boldsymbol{x}_i)\right]\left[m(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right]\left[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\right]^T\boldsymbol{\theta}_j$, where $(2A_i - 1)$ is a Rademacher sequence, and independent of $(\boldsymbol{x}_i, \epsilon_i)$. Hence given $\{(\boldsymbol{x}_i, \epsilon_i)\}_{i=1}^n$, on the event $\mathcal{H}_n$, we have $\sup_{\boldsymbol{\beta}\in\mathbb{B}_1, m\in\mathbb{M}} n^{-1}\sum_{i=1}^n \left|\boldsymbol{\theta}_j^T\boldsymbol{\gamma}_i(\boldsymbol{\beta}, m)\right|^2 \leqslant Ch^2$, for some positive constant C, and any j. Therefore, by Massart's concentration inequality (e.g., Theorem 14.2, Bühlmann and van de Geer [2011]), given $\{(\boldsymbol{x}_i, \epsilon_i)\}_{i=1}^n$, $\forall\, t > 0$,

$$P\Bigg(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1, m\in\mathbb{M}}\left|(nh)^{-1}\sum_{i=1}^n \boldsymbol{\theta}_j^T\boldsymbol{\gamma}_i(\boldsymbol{\beta}, m)\right| \geqslant (\sqrt{n}h)^{-1}\mathrm{E}_{sup}\big[\boldsymbol{\theta}_j^T\boldsymbol{\gamma}_i(\boldsymbol{\beta}, m)\big] + t\ \Bigg|\{(\boldsymbol{x}_i, \epsilon_i)\}_{i=1}^n, \mathcal{H}_n\cap\mathcal{J}_n\cap\mathcal{K}_n\Bigg)$$
$$\leqslant \exp\left(-\frac{nt^2}{8}\right),$$

where $\mathrm{E}_{sup}\big[\boldsymbol{\theta}_j^T\boldsymbol{\gamma}_i(\boldsymbol{\beta}, m)\big] = \mathrm{E}\left[\sup_{\boldsymbol{\beta}\in\mathbb{B}_1, m\in\mathbb{M}}\left|n^{-1/2}\sum_{i=1}^n \boldsymbol{\theta}_j^T\boldsymbol{\gamma}_i(\boldsymbol{\beta}, m)\right|\ \Big|\{(\boldsymbol{x}_i, \epsilon_i)\}_{i=1}^n, \mathcal{H}_n\cap\mathcal{J}_n\cap\mathcal{K}_n\right]$. Equivalently, $\forall\, t > 0$,

$$P\Bigg(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1, m\in\mathbb{M}}\left|n^{-1/2}\sum_{i=1}^n \boldsymbol{\theta}_j^T\boldsymbol{\gamma}_i(\boldsymbol{\beta}, m)\right| \geqslant \mathrm{E}_{sup}\big[\boldsymbol{\theta}_j^T\boldsymbol{\gamma}_i(\boldsymbol{\beta}, m)\big] + \sqrt{n}ht\ \Bigg|\{(\boldsymbol{x}_i, \epsilon_i)\}_{i=1}^n, \mathcal{H}_n\cap\mathcal{J}_n\cap\mathcal{K}_n\Bigg)$$
$$\leqslant \exp\left(-\frac{nt^2}{8}\right). \tag{S34}$$

Next we will derive an upper bound for $\mathrm{E}_{sup}\big[\boldsymbol{\theta}_j^T\boldsymbol{\gamma}_i(\boldsymbol{\beta}, m)\big]$. Let $M_1, \cdots, M_{m(s)}$ denote all possible subsets of $\{1, \cdots, p\}$, corresponding to different submodels of sizes at most ks. Note that $m(s) \leqslant \binom{p}{ks}$. Let $S_{M_l} = \{\boldsymbol{\beta}\in\mathbb{B}_1 : \mathrm{supp}(\boldsymbol{\beta}) = M_l\}$, where $\mathrm{supp}(\boldsymbol{\beta})$ denotes the support set of $\boldsymbol{\beta}$. Then $\mathbb{B}_1 = \bigcup_{l=1}^{m(s)} S_{M_l}$.

Given any $\delta_n > 0$, (w.l.o.g., $\delta_n \leqslant c\sqrt{s}h^2$), for each S_{M_l}, $l = 1, \cdots, m(s)$, we can cover it by L_2-balls of radius δ_n. Note that this cover has cardinality

$$N_l \leqslant \left(1 + \frac{2c_0\sqrt{s}h^2}{\delta_n}\right)^{ks},$$

for some positive constant c_0. Denote the centers of these L_2-balls by $\boldsymbol{\beta}_{l0}^{\circ}, \cdots, \boldsymbol{\beta}_{lN_l}^{\circ}$. Denote the collection of these L_2-balls by $\mathbb{C}(\boldsymbol{\beta}_{ll'}^{\circ})$, $l = 1, \cdots, m(s)$, $l' = 1, \cdots, N_l$.

Observe that on the event $\mathcal{J}_n$, $\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}|\boldsymbol{x}_i^T\boldsymbol{\beta}| \in T$, then $\bigcup_{\boldsymbol{\beta}\in\mathbb{B}_1} m(\cdot|\boldsymbol{\beta}) \in C_1^1(T)$. By Theorem 2.7.1 in van der Vaart and Wellner [1996], the entropy of the δ_n-covering number of $C_1^1(T)$ satisfies

$$\log N\left(\delta_n, C_1^1(T), L_2(\mathbb{P}_n)\right) \leqslant C||\boldsymbol{\beta}_0||_2 \frac{\sqrt{\log(p \vee n)}}{\delta_n},$$

for some positive constant C. So we can cover $C_1^1(T)$ with $N_2 \leqslant \exp\left[C||\boldsymbol{\beta}_0||_2 \delta_n^{-1}\sqrt{\log(p \vee n)}\right]$ L_2-balls of radius δ_n. Let the centers of these $L_2(\mathbb{P}_n)-$balls of the cover be $m_a^\circ(\cdot)$, $a = 1, \cdots, N_2$. Hence $\forall\ \boldsymbol{\beta} \in \mathbb{B}_1$, $m(\cdot|\boldsymbol{\beta}) \in C_1^1(T)$, we can find $l,\ l'$, and a function $m_a^\circ(\cdot) : T \to \mathbb{R}$ such that $\boldsymbol{\beta} \in \mathbb{C}(\boldsymbol{\beta}_{ll'}^\circ)$, and

$$n^{-1}\sum_{i=1}^n [m(\boldsymbol{x}_i^T\boldsymbol{\beta}_1|\boldsymbol{\beta}) - m_a^\circ(\boldsymbol{x}_i^T\boldsymbol{\beta}_1)]^2 \leqslant \delta_n^2, \quad \forall \boldsymbol{\beta}_1 \in \mathbb{B}_1. \tag{S35}$$

On the event $\mathcal{H}_n \cap \mathcal{J}_n \cap \mathcal{K}_n$, there exist some positive constants c_0, c_1 and c_2, such that

$$\begin{aligned}
& n^{-1}\left|\sum_{i=1}^n \boldsymbol{\theta}_j^T\boldsymbol{\gamma}_i(\boldsymbol{\beta}, m) - \sum_{i=1}^n \boldsymbol{\theta}_j^T\boldsymbol{\gamma}_i(\boldsymbol{\beta}_{ll'}^\circ, m_a^\circ)\right| \\
\leqslant & n^{-1}\left|\sum_{i=1}^n \left[m(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - m_a^\circ(\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^\circ)\right]\widetilde{\epsilon}_i[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^T\boldsymbol{\theta}_j\right| \\
\leqslant & \sqrt{n^{-1}\sum_{i=1}^n \left[m(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - m_a^\circ(\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^\circ)\right]^2} * \sqrt{n^{-1}\sum_{i=1}^n \{\widetilde{\epsilon}_i[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^T\boldsymbol{\theta}_j\}^2} \\
\leqslant & c_0\sqrt{n^{-1}\sum_{i=1}^n \left[m(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - m(\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^\circ|\boldsymbol{\beta})\right]^2 + n^{-1}\sum_{i=1}^n \left[m(\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^\circ|\boldsymbol{\beta}) - m_a^\circ(\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^\circ)\right]^2} \\
\leqslant & c_1\sqrt{n^{-1}\sum_{i=1}^n \left[\boldsymbol{x}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}_{ll'}^\circ)\right]^2 + \delta_n^2} \\
\leqslant & c_2\delta_n,
\end{aligned}$$

where the third inequality applies the event $\mathcal{H}_n$; the second last inequality applies (S35) and the differentiability condition; the last inequality applies the event $\mathcal{K}_n$. Hence, the

δ_n−covering number of the class of functions $\Gamma_j = \{\boldsymbol{\theta}_j^T \boldsymbol{\gamma}(Z, \boldsymbol{\beta}, m) : \boldsymbol{\beta} \in \mathbb{B}_1, m \in \mathbb{M}\}$ satisfies

$$N(\delta_n, \Gamma_j, L_1(\mathbb{P}_n)) \leqslant c\binom{p}{ks}\left(1 + \frac{2c_0\sqrt{s}h^2}{\delta_n}\right)^{ks} \exp\left[C\delta_n^{-1}\sqrt{\log(p \vee n)}\right], \quad \forall j. \tag{S36}$$

Recall that $\boldsymbol{\theta}_j^T \boldsymbol{\gamma}_i(\boldsymbol{\beta}, m) = 2(2A_i - 1)\left[\epsilon_i + g(\boldsymbol{x}_i)\right]\left[m(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right]\left[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\right]^T\boldsymbol{\theta}_j$, where $(2A_i - 1)$ is a Rademacher sequence independent of $(\boldsymbol{x}_i, \epsilon_i)$. Note that on the event $\mathcal{H}_n \cap \mathcal{J}_n \cap \mathcal{K}_n$, we have that $\sup_{\boldsymbol{\beta}\in\mathbb{B}_1, m\in\mathbb{M}} \sqrt{n^{-1}\sum_{i=1}^n \left[\boldsymbol{\theta}_j^T\boldsymbol{\gamma}_i(\boldsymbol{\beta}, m)^2\right]} = c_0 h \triangleq R_n$. Let $L = \min\{l : l \geqslant 1,\ 2^{-l} \leqslant 4/\sqrt{n}\}$. Therefore, Lemma 14.18 in van der Vaart and Wellner [1996] implies that

$$\begin{aligned}
&\mathrm{E}_{sup}\left[\boldsymbol{\theta}_j^T\boldsymbol{\gamma}_i(\boldsymbol{\beta}, m)\right] \\
=&\mathrm{E}\left[\sup_{\boldsymbol{\beta}\in\mathbb{B}_1, m\in\mathbb{M}}\left|n^{-1/2}\sum_{i=1}^n \boldsymbol{\theta}_j^T\boldsymbol{\gamma}_i(\boldsymbol{\beta}, m)\right| \,\middle|\, \{(\boldsymbol{x}_i, \epsilon_i)\}_{i=1}^n, \mathcal{H}_n \cap \mathcal{J}_n \cap \mathcal{K}_n\right] \\
\leqslant& Ch\left\{4 + 6\sum_{l=1}^L 2^{-l}\sqrt{\log\left[N(2^{-l}h, \Gamma_j, L_1(\mathbb{P}_n)) + 1\right]}\right\} \\
\leqslant& Ch\left\{4 + c_1\sum_{l=1}^L 2^{-l}\sqrt{ks\log p + ks\log\left[1 + \frac{2c_0\sqrt{s}h^2}{2^{-l}h}\right] + \frac{\sqrt{\log(p \vee n)}}{2^{-l}h}}\right\},
\end{aligned}$$

for some positive constants c_0, c_1, where the last inequality applies (S36). Hence we have $2^{-l} \geqslant 2/\sqrt{n}$, for any $1 \leqslant l \leqslant L$. Note that

$$\frac{2\sqrt{s}h^2}{2^{-l}h} \leqslant 4h\sqrt{ns} \leqslant 4c_2\sqrt{n^2h^7} \leqslant 4c_2 n,$$

for some positive constant c_2, by the assumptions of Theorem 1. Furthermore, the assumptions of Theorem 1 imply $s\sqrt{\log(p \vee n)} \leqslant c_1 n h^5 \leqslant c_2 h^{-1}$, for some positive constants c_1, c_2.

We thus have

$$
\begin{aligned}
&\mathrm{E}_{sup}\big[\boldsymbol{\theta}_j^T\boldsymbol{\gamma}_i(\boldsymbol{\beta},m)\big]\\
\leqslant& Ch\left[8+d_1\sum_{l=1}^{L}2^{-l}\sqrt{s\log(p\vee n)+2^l h^{-1}\sqrt{\log(p\vee n)}}\right]\\
\leqslant& Ch\left[8+d_1\sum_{l=1}^{L}2^{-l}\sqrt{c_2h^{-1}\sqrt{\log(p\vee n)}+2^l h^{-1}\sqrt{\log(p\vee n)}}\right]\\
\leqslant& Ch\left[8+d_2\big[h^{-1}\sqrt{\log(p\vee n)}\big]^{1/2}\sum_{l=1}^{L}2^{-l}(1+2^{l/2})\right]\\
\leqslant& Ch\left[8+2d_2\big[h^{-1}\sqrt{\log(p\vee n)}\big]^{1/2}\right]\\
\leqslant& d_3\left[h^2\log(p\vee n)\right]^{1/4},
\end{aligned}
$$

for some positive constants d_1, d_2, d_3, and all n sufficiently large, since $s\log(p\vee n)\leqslant s\sqrt{\log(p\vee n)}*\sqrt{\log(p\vee n)}\leqslant c_1h^{-1}\sqrt{\log(p\vee n)}$, and $\big[h^{-1}\sqrt{\log(p\vee n)}\big]^{1/2}\geqslant 8$ for some positive constant c_1, and all n sufficiently large. It follows from (S34) that $\forall\ t>0$,

$$
\begin{aligned}
&P\left(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1,m\in\mathbb{M}}\left|n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\theta}_j^T\boldsymbol{\gamma}_i(\boldsymbol{\beta},m)\right|\geqslant c\left[h^2\log(p\vee n)\right]^{1/4}+\sqrt{n}ht\ \middle|\{(\boldsymbol{x}_i,\epsilon_i)\}_{i=1}^n,\mathcal{H}_n\cap\mathcal{J}_n\cap\mathcal{K}_n\right)\\
\leqslant& \exp\left(-\frac{nt^2}{8}\right).
\end{aligned}
$$

Take $t=4\sqrt{n^{-1}\log(p\vee n)}$. Note that the assumptions of Theorem 1 imply $h\sqrt{\log(p\vee n)}\leqslant c_1\sqrt{nh^7}\leqslant c_1$, for some positive constant c_1. Hence we have

$$
\begin{aligned}
&P\left(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1,m\in\mathbb{M}}\left|n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\theta}_j^T\boldsymbol{\gamma}_i(\boldsymbol{\beta},m)\right|\geqslant c\left[h^2\log(p\vee n)\right]^{1/4}\ \middle|\mathcal{H}_n\cap\mathcal{J}_n\cap\mathcal{K}_n\right)\\
=&\mathrm{E}_{(\boldsymbol{x}_i,\epsilon_i)}\left\{P\left(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1,m\in\mathbb{M}}\left|n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\theta}_j^T\boldsymbol{\gamma}_i(\boldsymbol{\beta},m)\right|\geqslant c\left[h^2\log(p\vee n)\right]^{1/4}\ \middle|\{(\boldsymbol{x}_i,\epsilon_i)\}_{i=1}^n,\mathcal{H}_n\cap\mathcal{J}_n\cap\mathcal{K}_n\right)\right\}\\
\leqslant& \exp\left[-2\log(p\vee n)\right].
\end{aligned}
$$

Therefore, there exist positive constants c_0, c_1, such that for all n sufficiently large,

$$P\left(\max_{2\leqslant j\leqslant p}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1,m\in\mathbb{M}}\left|n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\theta}_j^T\boldsymbol{\gamma}_i(\boldsymbol{\beta},m)\right|\geqslant c_0\left[h^2\log(p\vee n)\right]^{1/4}\right)\leqslant \exp[-c_1\log(p\vee n)].$$

$\square$

Proof of Lemma A9. We will prove (S9) and (S11) below. The proofs of (S10) and (S12) are similar.

Recall that $\widetilde{\boldsymbol{x}}_i=\boldsymbol{x}_i-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)$, and $\widehat{\boldsymbol{x}}_i=\boldsymbol{x}_i-\widehat{\mathrm{E}}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})$. To prove (S9), observe that

$$\begin{aligned}
&\max_{2\leqslant j\leqslant p}\left|n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\theta}_j^T\boldsymbol{\nu}_1(Z_i,\widehat{\boldsymbol{\beta}},\widehat{G}^{(1)})\right|\\
&\leqslant \max_{2\leqslant j\leqslant p}\left|n^{-1/2}\sum_{i=1}^{n}\left[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)-G(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})-G^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\widetilde{\boldsymbol{x}}_{i,-1}^T(\boldsymbol{\beta}_{0,-1}-\widehat{\boldsymbol{\beta}}_{-1})\right]G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\right|\\
&\quad+\max_{2\leqslant j\leqslant p}\left|n^{-1/2}\sum_{i=1}^{n}\left[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)-G(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})-G^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\widetilde{\boldsymbol{x}}_{i,-1}^T(\boldsymbol{\beta}_{0,-1}-\widehat{\boldsymbol{\beta}}_{-1})\right]\right.\\
&\qquad\qquad\left.*\left[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})-G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right]\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\right|\\
&\quad+\max_{2\leqslant j\leqslant p}\left|n^{-1/2}\sum_{i=1}^{n}\left[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)-G(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\right]\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)(\widehat{\boldsymbol{x}}_{i,-1}-\widetilde{\boldsymbol{x}}_{i,-1})^T\boldsymbol{\theta}_j\right|\\
&\quad+\max_{2\leqslant j\leqslant p}\left|n^{-1/2}\sum_{i=1}^{n}G^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)(\boldsymbol{\beta}_{0,-1}-\widehat{\boldsymbol{\beta}}_{-1})^T(\widehat{\boldsymbol{x}}_{i,-1}\widehat{\boldsymbol{x}}_{i,-1}^T-\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T)\boldsymbol{\theta}_j\right|\\
&=\sum_{k=1}^{4}V_{n1k},
\end{aligned}$$

where the definition of V_{n1k}, $k=1,\cdots,4$, is clear from the context.

We first bound V_{n13}. Let $\mathcal{R}_n=\left\{b/2\leqslant\max_{1\leqslant i\leqslant n}\left|\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right|\leqslant 2b\right\}$. Note that $\max_{1\leqslant i\leqslant n}\left|G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right|\leqslant b$ by Assumption (A1)-(b). Lemma A6 implies that $P(\mathcal{R}_n)\geqslant 1-\exp[-c\log(p\vee n)]$ for some positive constant c, and all n sufficiently large. Then there exist positive constants c_0, c_1, c_2, c_3, such that for all n sufficiently large,

$$
\begin{aligned}
V_{n13} \leqslant & 2b\sqrt{n}\sqrt{n^{-1}\sum_{i=1}^{n}\Big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) - G(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\Big]^2} * \max_{2\leqslant j\leqslant p}\sqrt{n^{-1}\sum_{i=1}^{n}\big[(\widehat{\boldsymbol{x}}_{i,-1} - \widetilde{\boldsymbol{x}}_{i,-1})^T\boldsymbol{\theta}_j\big]^2} \\
\leqslant & c_0\sqrt{n}||\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}||_2 * sh^2\sqrt{\log(p \vee n)} \\
\leqslant & c_1\sqrt{ns}h^3 * sh\sqrt{\log(p \vee n)} \leqslant c_2\sqrt{ns}h^3,
\end{aligned}
$$

with probability at least $1 - \exp(-c_3 \log p)$. In the above, the second inequality applies (S6) in Lemma A4, Lemma A11 and Lemma A13, the third and fourth inequalities apply Theorem 1 and its assumptions.

We next bound V_{n14}. Similarly as in the proof of Lemma B10, we can show that

$$
\left\|n^{-1}\sum_{i=1}^{n}G^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)(\widehat{\boldsymbol{x}}_{i,-1}\widehat{\boldsymbol{x}}_{i,-1}^T - \widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T)\boldsymbol{\theta}_j\right\|_\infty \leqslant c_0\sqrt{s}h^2,
$$

with probability at least $1 - \exp(-c_1 \log p)$, for some positive constants c_0, c_1 and all n sufficiently large. Hence Theorem 1 and its assumptions imply that $V_{n14} \leqslant c_0\sqrt{n}s^{3/2}h^4 \leqslant d_0\sqrt{ns}h^4 * nh^5 \leqslant d_1\sqrt{ns}h^3$, with probability at least $1 - \exp(-c_1 \log p)$, for some positive constants c_0, c_1, d_0, d_1 and all n sufficiently large.

To bound V_{n12}, we note that

$$
\begin{aligned}
V_{n12} \leqslant & \max_{2\leqslant j\leqslant p}\left|n^{-1/2}\sum_{i=1}^{n}\Big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) - G(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\Big]\Big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}}) - G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\Big]\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\right| \\
& + \max_{2\leqslant j\leqslant p}\left|n^{-1/2}\sum_{i=1}^{n}G^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\Big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}}) - G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\Big](\boldsymbol{\beta}_{0,-1} - \widehat{\boldsymbol{\beta}}_{-1})^T\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\right| \\
= & V_{n121} + V_{n122},
\end{aligned}
$$

where the definitions of V_{n121} and V_{n122} are clear from the context. For V_{n121}, we have

$$
\begin{aligned}
V_{n121} \leqslant & \max_{1\leqslant i\leqslant n}\left|\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}}) - G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right| * \sqrt{n^{-1/2}\sum_{i=1}^n\Big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) - G(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\Big]^2} \\
& * \max_{2\leqslant j\leqslant p}\sqrt{n^{-1/2}\sum_{i=1}^n(\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j)^2} \\
\leqslant & c_0 h * ||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_2 \leqslant c_1\sqrt{ns}h^3,
\end{aligned}
$$

with probability at least $1-\exp(-c_2\log p)$, for some positive constants c_0, c_1, c_2 and all n sufficiently large. In the above, the second inequality applies Lemma A3, Lemma A4 and Lemma A11; the third inequality applies Theorem 1. Similarly, Theorem 1 and Lemma A11 imply that

$$
\begin{aligned}
V_{n122} \leqslant & \max_{1\leqslant i\leqslant n} b\left|\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}}) - G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right| * ||\widehat{\boldsymbol{\beta}}_{-1} - \boldsymbol{\beta}_{0,-1}||_1 * \max_{2\leqslant j\leqslant p}\left|\left|n^{-1/2}\sum_{i=1}^n\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\right|\right|_\infty \\
\leqslant & c_0 h * ||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_1 \leqslant c_1 s\sqrt{n}h^3,
\end{aligned}
$$

with probability at least $1-\exp(-c_2\log p)$, for some positive constants c_0, c_1, c_2 and all n sufficiently large. In the above, the second inequality applies Assumption (A5)-(a), Lemma A3 and Lemma B1; the third inequality applies Theorem 1.

Finally we bound V_{n11}. Note that $\forall\boldsymbol{\beta}\in\mathbb{B}_1$, $G^{(1)}(\cdot|\boldsymbol{\beta})$ is differentiable by Assumption (A5)-(a). Hence, there exists some positive constant L such that

$$
\left|G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\widehat{\boldsymbol{\beta}}) - G(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}}) - G^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\boldsymbol{x}_{i,-1}^T(\boldsymbol{\beta}_{0,-1} - \widehat{\boldsymbol{\beta}}_{-1})\right| \leqslant L\big[\boldsymbol{x}_i^T(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}})\big]^2.
$$

It implies that

$$
\begin{aligned}
V_{n11} \leqslant & \max_{2\leqslant j\leqslant p}\left|n^{-1/2}\sum_{i=1}^{n}\Big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)-G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\widehat{\boldsymbol{\beta}})\Big]G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\right| \\
& + \max_{2\leqslant j\leqslant p} L\left|n^{-1/2}\sum_{i=1}^{n}\big[\boldsymbol{x}_i^T(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}})\big]^2G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\right| \\
& + \max_{2\leqslant j\leqslant p}\left|n^{-1/2}\sum_{i=1}^{n}\big[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\big]^2(\boldsymbol{\beta}_{0,-1}-\widehat{\boldsymbol{\beta}}_{-1})^T\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\right| \\
& + \max_{2\leqslant j\leqslant p}\Bigg|n^{-1/2}\sum_{i=1}^{n}\big[G^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})-G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\big]G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) \\
& \qquad\qquad * (\boldsymbol{\beta}_{0,-1}-\widehat{\boldsymbol{\beta}}_{-1})^T\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\Bigg| \\
= & \sum_{l=1}^{4}V_{n11l},
\end{aligned}
$$

where the definition of V_{n11l}, $l=1,\cdots,4$, is clear from the context. To bound V_{n112}, note that

$$
V_{n112} \leqslant Lb\sqrt{n}\sqrt{n^{-1}\sum_{i=1}^{n}\big[\boldsymbol{x}_i^T(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}})\big]^4} * \max_{2\leqslant j\leqslant p}\sqrt{n^{-1}\sum_{i=1}^{n}(\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j)^2}.
$$

Note that $\frac{\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}}{||\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}||_2}\in\mathbb{K}(p,2ks)$ for $\widehat{\boldsymbol{\beta}}\neq\boldsymbol{\beta}_0$, where $\mathbb{K}(p,s_0)=\{\boldsymbol{v}\in\mathbb{R}^p:||\boldsymbol{v}||_2\leqslant 1,||\boldsymbol{v}||_0\leqslant s_0\}$. In Lemma B3, take $t=d_0\sigma_x^4||\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}||_2^2$ and $s_0=ks$, for some positive constant d_0, then Theorem 1 and Lemma B3 imply that $n^{-1}\sum_{i=1}^{n}\big[\boldsymbol{x}_i^T(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}})\big]^4\leqslant c_0\sigma_x^4s^2h^8$ with probability at least $1-\exp(-c_1\log p)$, for some positive constants c_0, c_1, and all n sufficiently large. We thus have $V_{n112}\leqslant d_0s\sqrt{n}h^4$ with probability at least $1-\exp[-d_1\log(p\vee n)]$, for some positive constants d_0, d_1, and all n sufficiently large.

To bound V_{n113}, note that $\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)$ and $\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j$ are both sub-Gaussian by Lemma A2, and $G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)$ is bounded by Assumption (A5)-(a). It's easy to show $G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\big]^2\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j$

is also sub-Gaussian by Lemma A2. We observe that

$$
\begin{aligned}
&\mathrm{E}\Big\{\big[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\big]^2\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\Big\}\\
=&\mathrm{E}_{\boldsymbol{x}_i^T\boldsymbol{\beta}_0}\Big\{\big[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\big]^2\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\mathrm{E}\big(\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\big|\boldsymbol{x}_i^T\boldsymbol{\beta}_0\big)\Big\}\\
=&\mathbf{0}_{p-1},
\end{aligned}
$$

where $\mathbf{0}_{p-1}$ is a $(p-1)-$dimensional vector with all entries 0, since $\mathrm{E}\big(\widetilde{\boldsymbol{x}}_{i,-1}\big|\boldsymbol{x}_i^T\boldsymbol{\beta}_0\big) = \mathbf{0}_{p-1}$. Hence, Lemma B1 implies that

$$
\begin{aligned}
V_{n113} \leqslant&\sqrt{n}||\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}||_1 * \max_{2\leqslant j\leqslant p}\bigg\|n^{-1}\sum_{i=1}^n\big[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\big]^2\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j\bigg\|_\infty\\
\leqslant& c_0\sqrt{n}sh^2 * \sqrt{\frac{\log p}{n}} \leqslant c_0 s\sqrt{n}h^{9/2},
\end{aligned}
$$

with probability at least $1-\exp(-c_1\log p)$, for some positive constants c_0, c_1, and all n sufficiently large, where the second inequality also applies Theorem 1, and the third inequality applies the assumptions of Theorem 1.

Theorem 1 and Assumption (A5)-(b) imply that

$$
V_{n114} \leqslant c_0\sqrt{n}||\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0||_2\Big\{n^{-1}\sum_{i=1}^n\big[(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}})^T\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^4\Big\}^{1/4} * \max_{2\leqslant j\leqslant p}\Big[n^{-1}\sum_{i=1}^n(\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j)^4\Big]^{1/4},
$$

with probability at least $1-\exp(-c_1\log p)$, for some positive constants c_0, c_1, and all n sufficiently large. Lemma B3 implies that $n^{-1}\sum_{i=1}^n\big[(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}})^T\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^4 \leqslant d_0||\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}||_2^4$, and $n^{-1}\sum_{i=1}^n(\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j)^4 \leqslant d_0$, with probability at least $1-\exp(-d_1\sqrt{n})$, for some positive constants d_0, d_1, and all n sufficiently large. Hence combining the results with Theorem 1, we have $V_{n114} \leqslant c_0\sqrt{n}sh^4$, with probability at least $1-\exp(-c_1\log p)$, for some positive constants c_0, c_1, and all n sufficiently large.

We finally bound V_{n111}. Theorem 1 implies that $P\left(\widehat{\boldsymbol{\beta}}\in\mathbb{B}_1\right) \geqslant 1-\exp(-c\log p)$, for some positive constant c, and all n sufficiently large. It is sufficient to show that there exist some positive constants c_0, c_1, such that for all n sufficiently large,

$$P\left(\max_{2\leqslant j\leqslant p}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\theta}_j^T\boldsymbol{\nu}_{11}(Z_i,\boldsymbol{\beta})\right|\geqslant c_0h^2s\sqrt{\log(p\vee n)}\right)\leqslant\exp[-c_1\log(p\vee n)],$$

where $Z_i=(\boldsymbol{x}_i,\epsilon_i,A_i)$, $\boldsymbol{\nu}_{11}(Z_i,\boldsymbol{\beta})=\big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)-G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta})\big]G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\widetilde{\boldsymbol{x}}_{i,-1}$. Note that the assumptions of Theorem 1 imply that $h^2s\sqrt{\log(p\vee n)}\leqslant d_0h^2\sqrt{snh^5}\leqslant d_0h^3\sqrt{sn}$ for some positive constant d_0. Note that for any $\boldsymbol{\beta}\in\mathbb{B}_1$, we have

$$\begin{aligned}\mathrm{E}\big[\boldsymbol{\theta}_j^T\boldsymbol{\nu}_{11}(Z_i,\boldsymbol{\beta})\big]=&\mathrm{E}_{\boldsymbol{x}_i^T\boldsymbol{\beta}_0}\mathrm{E}\Big\{\big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)-G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta})\big]G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j|\boldsymbol{x}_i^T\boldsymbol{\beta}_0\Big\}\\ =&\mathrm{E}_{\boldsymbol{x}_i^T\boldsymbol{\beta}_0}\Big\{\big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)-G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta})\big]G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\boldsymbol{\theta}_j^T\mathrm{E}(\widetilde{\boldsymbol{x}}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\Big\}\\ =&0.\end{aligned}$$

Denote the event

$$\mathcal{T}_n=\Big\{\max_{2\leqslant j\leqslant p}\frac{1}{n}\sum_{i=1}^{n}(\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j)^4\leqslant 20\xi_2^{-4}\sigma_x^4,\text{ and }\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\frac{1}{n}\sum_{i=1}^{n}\big[\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\big]^4\leqslant 5\sigma_x^4s^2h^8\Big\}.$$

Similarly as the above analysis, Lemma A2 and Lemma B3 imply that $P(\mathcal{T}_n)\geqslant 1-\exp(-d_0\sqrt{n})$, for some positive constant d_0 and all n sufficiently large.

To prove (S39), we have

$$\begin{aligned}&P\left(\max_{2\leqslant j\leqslant p}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\theta}_j^T\boldsymbol{\nu}_{11}(Z_i,\boldsymbol{\beta})\right|>t\right)\\ \leqslant&P\left(\max_{2\leqslant j\leqslant p}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\theta}_j^T\boldsymbol{\nu}_{11}(Z_i,\boldsymbol{\beta})\right|>t\;\Big|\mathcal{G}_n\cap\mathcal{K}_n\cap\mathcal{T}_n\right)+\exp[-c\log(p\vee n)]\\ \leqslant&\sum_{j=2}^{p}P\left(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\theta}_j^T\boldsymbol{\nu}_{11}(Z_i,\boldsymbol{\beta})\right|>t\;\Big|\mathcal{G}_n\cap\mathcal{K}_n\cap\mathcal{T}_n\right)+\exp[-c\log(p\vee n)],\end{aligned}$$

for some positive constant c, and all n sufficiently large, where the events $\mathcal{G}_n$ and $\mathcal{K}_n$ are defined in Lemma A3.

We observe that there exists some $\boldsymbol{\beta}^r$ between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$, such that

$$
\begin{aligned}
&\sup_{\boldsymbol{\beta}\in\mathbb{B}_1} \frac{1}{n}\sum_{i=1}^n [\boldsymbol{\theta}_j^T \boldsymbol{\nu}_{11}(Z_i,\boldsymbol{\beta})]^2 \\
\leqslant& b^2 \sup_{\boldsymbol{\beta}\in\mathbb{B}_1} \frac{1}{n}\sum_{i=1}^n \big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) - G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta})\big]^2 (\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j)^2 \\
\leqslant& b^2 \sup_{\boldsymbol{\beta}\in\mathbb{B}_1} \frac{1}{n}\sum_{i=1}^n \Big\{ - f_0'(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})^T(\boldsymbol{\beta}_{-1}-\boldsymbol{\beta}_{0,-1}) \\
&\qquad\qquad + \frac{1}{2}\mathrm{E}\big\{f_0''(\boldsymbol{x}_i^T\boldsymbol{\beta}^r)[\boldsymbol{x}_{i,-1}^T(\boldsymbol{\beta}_{-1}-\boldsymbol{\beta}_{0,-1})]^2|\boldsymbol{x}_i^T\boldsymbol{\beta}\big\}\Big\}^2 (\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j)^2 \\
\leqslant& 2b^2 \sup_{\boldsymbol{\beta}\in\mathbb{B}_1} \frac{1}{n}\sum_{i=1}^n \big[f_0'(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})^T(\boldsymbol{\beta}_{-1}-\boldsymbol{\beta}_{0,-1})\big]^2 (\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j)^2 \\
&+ c_0 b^2 \sup_{\boldsymbol{\beta}\in\mathbb{B}_1} \frac{1}{n}\sum_{i=1}^n \big[(\boldsymbol{\beta}_{-1}-\boldsymbol{\beta}_{0,-1})^T\mathrm{E}(\boldsymbol{x}_{i,-1}\boldsymbol{x}_{i,-1}^T|\boldsymbol{x}_i^T\boldsymbol{\beta})^T(\boldsymbol{\beta}_{-1}-\boldsymbol{\beta}_{0,-1})\big](\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j)^2 \\
\leqslant& 2b^2 \sup_{\boldsymbol{\beta}\in\mathbb{B}_1} \frac{1}{n}\sum_{i=1}^n \big[f_0'(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\big]^2 (\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j)^2 \\
&+ c_0 b^2 \sup_{\boldsymbol{\beta}\in\mathbb{B}_1} \sqrt{\frac{1}{n}\sum_{i=1}^n \big[\lambda_{\max}(\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i^T|\boldsymbol{x}_i^T\boldsymbol{\beta}))\big]^2 ||\boldsymbol{\beta}-\boldsymbol{\beta}_0||_2^4}\sqrt{\frac{1}{n}\sum_{i=1}^n (\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j)^4} \\
\leqslant& b^4 \sup_{\boldsymbol{\beta}\in\mathbb{B}_1} \frac{1}{n}\sum_{i=1}^n \big[\mathrm{E}\big(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}\big)^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\big]^2 (\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j)^2 + c_0 b^2\sqrt{\xi_4}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}||\boldsymbol{\beta}-\boldsymbol{\beta}_0||_2^2\sqrt{\frac{1}{n}\sum_{i=1}^n (\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j)^4},
\end{aligned}
$$

with probability at least $1-\exp[-c_2\log(p\vee n)]$, for some positive constants c_0, c_1, c_2, and all n sufficiently large. In the above, the second inequality applies (S25) in the proof of Lemma A4; the third inequality applies Assumption (A1)-(b); the last inequality applies Assumption (A2)-(a) and (A1)-(b). On the event $\mathcal{T}_n$, we have

$$
\begin{aligned}
&\sup_{\boldsymbol{\beta}\in\mathbb{B}_1} \frac{1}{n}\sum_{i=1}^n \big[\mathrm{E}\big(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}\big)^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\big]^2 (\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j)^2 \\
\leqslant& \sqrt{\sup_{\boldsymbol{\beta}\in\mathbb{B}_1} \frac{1}{n}\sum_{i=1}^n \big[\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\big]^4}\sqrt{\frac{1}{n}\sum_{i=1}^n (\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j)^4} \\
\leqslant& d_0 s h^4,
\end{aligned}
$$

for some positive constant d_0. Hence we obtain that $\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\frac{1}{n}\sum_{i=1}^n[\boldsymbol{\theta}_j^T\boldsymbol{\nu}_{11}(Z_i,\boldsymbol{\beta})]^2 \leqslant d_1 s h^4$, on the event $\mathcal{T}_n$, with probability at least $1-\exp[-d_2\log(p\vee n)]$, for some positive constants d_1, d_2, and all n sufficiently large. Therefore, by Massart's concentration inequality (e.g., Theorem 14.2, Bühlmann and van de Geer [2011]) on the event $\mathcal{G}_n\cap\mathcal{T}_n\cap\mathcal{K}_n$, $\forall\ t>0$,

$$P\left(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|n^{-1/2}\sum_{i=1}^n\boldsymbol{\theta}_j^T\boldsymbol{\nu}_{11}(Z_i,\boldsymbol{\beta})\right|\geqslant E_{sup,1}+th^2\sqrt{sn}\middle|\mathcal{G}_n\cap\mathcal{K}_n\cap\mathcal{T}_n\right)\leqslant\exp\left(-\frac{nt^2}{8}\right),\quad\text{(S37)}$$

where $E_{sup,1}=\mathrm{E}\left[\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|n^{-1/2}\sum_{i=1}^n\boldsymbol{\theta}_j^T\boldsymbol{\nu}_{11}(Z_i,\boldsymbol{\beta})\right|\middle|\mathcal{G}_n\cap\mathcal{K}_n\cap\mathcal{T}_n\right]$.

Similarly as the proof of Lemma A8, we can cover $\mathbb{B}_1$ by $\binom{p}{ks}\left(1+\frac{2c_0sh^2\sqrt{\log(p\vee n)}}{\delta_n}\right)^{ks}$ L_2-balls of radius $\frac{\delta_n}{\sqrt{s\log(p\vee n)}}$. Denote the centers of these L_2-balls by $\boldsymbol{\beta}_{l0}^{\circ},\cdots,\boldsymbol{\beta}_{lN_l}^{\circ}$. Denote these L_2-balls by $\mathbb{C}(\boldsymbol{\beta}_{ll'}^{\circ})$, $l=1,\cdots,m(s)$, $l'=1,\cdots,N_l$. Hence $\forall\ \boldsymbol{\beta}\in\mathbb{B}_1$, we can find $l,\ l'$ such that $\boldsymbol{\beta}\in\mathbb{C}(\boldsymbol{\beta}_{ll'}^{\circ})$. On the event $\mathcal{G}_n\cap\mathcal{K}_n$, there exist some positive constants c_0, c_1, such that for all n sufficiently large,

$$\begin{aligned}&n^{-1}\left|\sum_{i=1}^n\boldsymbol{\theta}_j^T\boldsymbol{\nu}_{11}(Z_i,\boldsymbol{\beta})-\sum_{i=1}^n\boldsymbol{\theta}_j^T\boldsymbol{\nu}_{11}(Z_i,\boldsymbol{\beta}_{ll'}^{\circ})\right|\\
\leqslant&n^{-1}\left|\sum_{i=1}^n\left[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta})-G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_{ll'}^{\circ})\right]G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)[\boldsymbol{x}_{i,-1}-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^T\boldsymbol{\theta}_j\right|\\
\leqslant&\sqrt{n^{-1}\sum_{i=1}^n[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta})-G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_{ll'}^{\circ})]^2}*\sqrt{n^{-1}\sum_{i=1}^n[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)[\boldsymbol{x}_{i,-1}-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^T\boldsymbol{\theta}_j]^2}\\
\leqslant&c_0||\boldsymbol{\beta}-\boldsymbol{\beta}_{ll'}^{\circ}||_2\sqrt{s\log(p\vee n)}=c_0\delta_n,\end{aligned}$$

with probability at least $1-\exp[-c_1\log(p\vee n)]$. In the above, the last inequality applies the event $\mathcal{G}_n\cap\mathcal{K}_n$, Assumption (A1)-(b) and Assumption (A5)-(c) Hence, the δ_n-covering number of the class of functions $V_{1j}=\{\boldsymbol{\theta}_j^T\boldsymbol{\nu}_{11}(Z_i,\boldsymbol{\beta}):\boldsymbol{\beta}\in\mathbb{B}_1\}$ is bounded by

$$N(\delta_n,V_{1j},L_1(\mathbb{P}_n))\leqslant c\binom{p}{ks}\left(1+\frac{2c_0sh^2\sqrt{\log(p\vee n)}}{\delta_n}\right)^{ks}.\quad\text{(S38)}$$

Let $a_1,\cdots,a_n$ be a Rademacher sequence that is independent of the data. The symmetrization theorem (Theorem 14.3 in Bühlmann and van de Geer [2011]) implies that

$$\begin{aligned}
E_{sup,1} =& \mathrm{E}\left[\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|n^{-1/2}\sum_{i=1}^n \boldsymbol{\theta}_j^T\boldsymbol{\nu}_{11}(Z_i,\boldsymbol{\beta})\right|\,\middle|\mathcal{G}_n\cap\mathcal{K}_n\cap\mathcal{T}_n\right]\\
\leqslant& 2\mathrm{E}_{\{\boldsymbol{x}_i\}_{i=1}^n}\left\{\mathrm{E}\left[\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|n^{-1/2}\sum_{i=1}^n a_i\boldsymbol{\theta}_j^T\boldsymbol{\nu}_{11}(Z_i,\boldsymbol{\beta})\right|\,\middle|\{\boldsymbol{x}_i\}_{i=1}^n,\mathcal{G}_n\cap\mathcal{K}_n\cap\mathcal{T}_n\right]\right\}.
\end{aligned}$$

Note that on the event $\mathcal{T}_n$, we have $\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\sqrt{n^{-1}\sum_{i=1}^n[a_i\boldsymbol{\theta}_j^T\boldsymbol{\nu}_{11}(Z_i,\boldsymbol{\beta})]^2}\leqslant ch^2\sqrt{s}\triangleq R_n$. Lemma 14.18 in van der Vaart and Wellner [1996] implies that

$$\begin{aligned}
E_{sup,1} \leqslant& 2\mathrm{E}_{\{\boldsymbol{x}_i\}_{i=1}^n}\left\{\mathrm{E}\left[\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|n^{-1/2}\sum_{i=1}^n a_i\boldsymbol{\theta}_j^T\boldsymbol{\nu}_{11}(Z_i,\boldsymbol{\beta})\right|\,\middle|\{\boldsymbol{x}_i\}_{i=1}^n,\mathcal{G}_n\cap\mathcal{K}_n\cap\mathcal{T}_n\right]\right\}\\
\leqslant& Ch^2\sqrt{s}\left\{4+6\sum_{l=1}^L 2^{-l}\sqrt{\log\left[N(2^{-l}h^2\sqrt{s},V_j,L_1(\mathbb{P}_n))+1\right]}\right\}\\
\leqslant& c_0h^2s\sqrt{\log(p\vee n)},
\end{aligned}$$

for some positive constants c_0, C, where $L=\min\{l: l\geqslant 1,\ 2^{-l}\leqslant 4/\sqrt{n}\}$, and the last inequality applies (S38). The analysis is similar as that in the proof of Lemma A8.

It follows from (S37) that

$$\begin{aligned}
&P\left(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|n^{-1/2}\sum_{i=1}^n \boldsymbol{\theta}_j^T\boldsymbol{\nu}_{11}(Z_i,\boldsymbol{\beta})\right|\geqslant c_0h^2s\sqrt{\log(p\vee n)}+th^2\sqrt{ns}\,\middle|\mathcal{G}_n\cap\mathcal{K}_n\cap\mathcal{T}_n\right)\\
\leqslant& \exp\left(-\frac{nt^2}{8}\right).
\end{aligned}$$

Take $t=4\sqrt{n^{-1}\log(p\vee n)}$. Then there exist some positive constants c_0, c_1, such that for all n sufficiently large,

$$P\left(\max_{2\leqslant j\leqslant p}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|n^{-1/2}\sum_{i=1}^n \boldsymbol{\theta}_j^T\boldsymbol{\nu}_{11}(Z_i,\boldsymbol{\beta})\right|\geqslant 4ch^2s\sqrt{\log(p\vee n)}\right)\leqslant \exp[-c_1\log(p\vee n)].$$

Combining all the above results, we prove (S9).

To prove (S11), we note that

$$
\begin{aligned}
&\max_{2\leqslant j\leqslant p}\left|n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\theta}_j^T\boldsymbol{\nu}_2(Z_i,\widehat{\boldsymbol{\beta}},\widehat{G},\widehat{G}^{(1)})\right|\\
&\leqslant \max_{2\leqslant j\leqslant p}\left|n^{-1/2}\sum_{i=1}^{n}\Big[G(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})-\widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\Big]G^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\big[\boldsymbol{x}_{i,-1}-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})\big]^T\boldsymbol{\theta}_j\right|\\
&\quad+\max_{2\leqslant j\leqslant p}\left|n^{-1/2}\sum_{i=1}^{n}\Big[G(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})-\widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\Big]\big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})-G^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\big]\big[\boldsymbol{x}_{i,-1}-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\theta}_j\right|\\
&\quad+\max_{2\leqslant j\leqslant p}\left|n^{-1/2}\sum_{i=1}^{n}\Big[G(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})-\widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\Big]\big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})-G^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\big]\right.\\
&\qquad\qquad\qquad\left.*\big[\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\theta}_j\right|\\
&\quad+\max_{2\leqslant j\leqslant p}\left|n^{-1/2}\sum_{i=1}^{n}\Big[G(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})-\widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\Big]\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\big[\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})-\widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})\big]^T\boldsymbol{\theta}_j\right|\\
&=\sum_{k=1}^{4}V_{n2k},
\end{aligned}
$$

where the definition of V_{n2k}, $k=1,\cdots,4$, is clear from the context.

We first bound V_{n22}. Lemma A5 and Lemma A6 imply that $\max_{1\leqslant i\leqslant n}\left|G(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})-\widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\right|\leqslant c_0h^2$, and $\max_{1\leqslant i\leqslant n}\left|\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})-G^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\right|\leqslant c_0h$, with probability at least $1-\exp(-c_1\log p)$, for some positive constants c_0, c_1 and all n sufficiently large. Note that $\boldsymbol{x}_{i,-1}-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)$ is sub-Gaussian by Lemma A2. Then Lemma A13 and Lemma B1 imply that $\max_{2\leqslant j\leqslant p}n^{-1}\sum_{i=1}^n\left\{\big[\boldsymbol{x}_{i,-1}-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\theta}_j\right\}^2\leqslant c_2$, with probability at least $1-\exp(-c_3n)$, for some positive constants c_2, c_3 and all n sufficiently large. Combining all these results, we have that $V_{n22}\leqslant d_0\sqrt{n}h^3$, with probability at least $1-\exp(-d_1\log p)$, for some positive constants d_0, d_1 and all n sufficiently large.

Next, Assumption (A2)-(c) implies that

$$\begin{aligned}
&\max_{2\leqslant j\leqslant p} n^{-1}\sum_{i=1}^{n}\Big\{\big[\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\theta}_j\Big\}^2\\
\leqslant&\max_{2\leqslant j\leqslant p} Cn^{-1}\sum_{i=1}^{n}\big[|\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}-\boldsymbol{x}_i^T\boldsymbol{\beta}_0|+(|\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|+|\boldsymbol{x}_i^T\boldsymbol{\beta}_0|)||\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0||_2\big]^2||\boldsymbol{\theta}_j||_2^2\\
\leqslant&\max_{2\leqslant j\leqslant p} 2C\xi_2^{-2}n^{-1}\sum_{i=1}^{n}\big[|\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}-\boldsymbol{x}_i^T\boldsymbol{\beta}_0|^2+(|\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|+|\boldsymbol{x}_i^T\boldsymbol{\beta}_0|)^2||\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0||_2^2\big]\\
\leqslant& c_0||\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0||_2^2,
\end{aligned}$$

with probability at least $1-\exp(-c_1 n)$, for some positive constants c_0, c_1 and all n sufficiently large. In the above, the second inequality applies Lemma A13; the last inequality applies Lemma B1. Combining Theorem 1 with the above results, we have $V_{n23}\leqslant d_0\sqrt{ns}h^5$, with probability at least $1-\exp(-d_1\log p)$, for some positive constants d_0, d_1 and all n sufficiently large.

Observe that Theorem 1, Lemma A7 and Lemma A13 imply that $\max_{1\leqslant i\leqslant n}\Big|\big[\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})-\widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})\big]^T\boldsymbol{\theta}_j\Big|\leqslant c_0h^2$, with probability at least $1-\exp(-c_1\log p)$, for some positive constants c_0, c_1 and all n sufficiently large. Then the above results and Assumption (A5)-(a) indicate that $V_{n24}\leqslant d_0\sqrt{n}h^4$, with probability at least $1-\exp(-d_1\log p)$, for some positive constants d_0, d_1 and all n sufficiently large.

Finally we bound V_{n21}. Define

$$\mathbb{U}=\Big\{u(\cdot|\boldsymbol{\beta}):\boldsymbol{\beta}\in\mathbb{B}_1,\ u(\cdot|\boldsymbol{\beta})\in C_1^1(T),\ \forall\ \boldsymbol{\beta},\ \text{and}\ \max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\big|u(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big|\leqslant c_1h^2\Big\},$$

for some positive constant c_1, where $T=\{t\in\mathbb{R}:|t|\leqslant 2||\boldsymbol{\beta}_0||_2\sigma_x\sqrt{\log(p\vee n)}\}$, and $C_1^1(T)$ is the set of all continuous and Lipschitz functions $f:T\to\mathbb{R}$. It is sufficient to show that there exist some positive constants c_0, c_1, such that for all n sufficiently large,

$$P\left(\max_{2\leqslant j\leqslant p}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1,u\in\mathbb{U}}\left|n^{-1/2}\sum_{i=1}^n\boldsymbol{\theta}_j^T\boldsymbol{\nu}_{21}(Z_i,\boldsymbol{\beta},u)\right|\geqslant c_0h[\log(p\vee n)]^{1/4}\right)\leqslant\exp[-c_1\log(p\vee n)],\tag{S39}$$

where $Z_i=(\boldsymbol{x}_i,\epsilon_i,A_i)$, $\boldsymbol{\nu}_{21}(Z_i,\boldsymbol{\beta},u)=\big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-u(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big[\boldsymbol{x}_{i,-1}-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})\big]$, $u(\boldsymbol{x}^T\boldsymbol{\beta}|\boldsymbol{\beta})$ depends on $\boldsymbol{x}$ only through $\boldsymbol{x}^T\boldsymbol{\beta}$. Note the assumptions of Theorem 1 imply that $h[\log(p\vee n)]^{1/4}\leqslant d_0(nh^9)^{1/4}\leqslant d_0h^3\sqrt{n}*(nh^3)^{-1/4}\leqslant d_0sh^3\sqrt{n}$ for some positive constant d_0.

To prove (S39), we have

$$\begin{aligned}&P\left(\max_{2\leqslant j\leqslant p}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1,u\in\mathbb{U}}\left|n^{-1/2}\sum_{i=1}^n\boldsymbol{\theta}_j^T\boldsymbol{\nu}_{21}(Z_i,\boldsymbol{\beta},u)\right|>t\right)\\ \leqslant&P\left(\max_{2\leqslant j\leqslant p}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1,u\in\mathbb{U}}\left|n^{-1/2}\sum_{i=1}^n\boldsymbol{\theta}_j^T\boldsymbol{\nu}_{21}(Z_i,\boldsymbol{\beta},u)\right|>t\;\middle|\mathcal{G}_n\cap\mathcal{J}_n\cap\mathcal{K}_n\right)+\exp[-c\log(p\vee n)]\\ \leqslant&\sum_{j=2}^pP\left(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1,u\in\mathbb{U}}\left|n^{-1/2}\sum_{i=1}^n\boldsymbol{\theta}_j^T\boldsymbol{\nu}_{21}(Z_i,\boldsymbol{\beta},u)\right|>t\;\middle|\mathcal{G}_n\cap\mathcal{J}_n\cap\mathcal{K}_n\right)+\exp[-c\log(p\vee n)],\end{aligned}$$

for some positive constant c, and all n sufficiently large, where the events $\mathcal{G}_n$, $\mathcal{J}_n$ and $\mathcal{K}_n$ are defined in Lemma A3. Note that $\mathrm{E}[\boldsymbol{\theta}_j^T\boldsymbol{\nu}_{21}(Z_i,\boldsymbol{\beta},u)]=0$, and $\sup\limits_{\boldsymbol{\beta}\in\mathbb{B}_1,u\in\mathbb{U}}n^{-1}\sum_{i=1}^n\left|\boldsymbol{\theta}_j^T\boldsymbol{\nu}_{21}(Z_i,\boldsymbol{\beta},u)\right|^2\leqslant Ch^4$, on the event $\mathcal{G}_n$, for some positive constant C, and any j. Hence by Massart's concentration inequality (e.g., Theorem 14.2, Bühlmann and van de Geer [2011]), $\forall\ t>0$,

$$P\Bigg(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1,u\in\mathbb{U}}\left|n^{-1/2}\sum_{i=1}^n\boldsymbol{\theta}_j^T\boldsymbol{\nu}_{21}(Z_i,\boldsymbol{\beta},u)\right|\geqslant E_{sup,2}+th^2\sqrt{n}\Bigg|\mathcal{G}_n\cap\mathcal{J}_n\cap\mathcal{K}_n\Bigg)\leqslant\exp\left(-\frac{nt^2}{8}\right),\tag{S40}$$

where $E_{sup,2}=\mathrm{E}\left[\sup_{\boldsymbol{\beta}\in\mathbb{B}_1,u\in\mathbb{U}}\left|n^{-1/2}\sum_{i=1}^n\boldsymbol{\theta}_j^T\boldsymbol{\nu}_{21}(Z_i,\boldsymbol{\beta},u)\right|\;\middle|\mathcal{G}_n\cap\mathcal{J}_n\cap\mathcal{K}_n\right]$.

Similarly as the proof of Lemma A8, we can cover $\mathbb{B}_1$ by $\binom{p}{ks}\left(1+\frac{2c_0s^{3/2}h^2\log(p\vee n)}{\delta_n^2}\right)^{ks}$ L_2-balls of radius $\frac{\delta_n^2}{s\log(p\vee n)}$. Denote the centers of these L_2-balls by $\boldsymbol{\beta}_{l0}^\circ,\cdots,\boldsymbol{\beta}_{lN_l}^\circ$. Denote these L_2-balls by $\mathbb{C}(\boldsymbol{\beta}_{ll'}^\circ)$, $l=1,\cdots,m(s)$, $l'=1,\cdots,N_l$.

Observe that on the event $\mathcal{J}_n$, $\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}|\boldsymbol{x}_i^T\boldsymbol{\beta}|\in T$, then $\bigcup_{\boldsymbol{\beta}\in\mathbb{B}_1}u(\cdot|\boldsymbol{\beta})\in C_1^1(T)$. By Theorem 2.7.1 in van der Vaart and Wellner [1996], we have

$$\log N\left(\delta_n, C_1^1(T), L_2(\mathbb{P}_n)\right)\leqslant C||\boldsymbol{\beta}_0||_2\frac{\sqrt{\log(p\vee n)}}{\delta_n},$$

for some positive constant C. So we can find $N_2\leqslant\exp\left[C||\boldsymbol{\beta}_0||_2\delta_n^{-1}\sqrt{\log(p\vee n)}\right]$ balls with the centers $u_a^\circ(\cdot)$, $a=1,\cdots,N_2$, to cover $C_1^1(T)$. Hence $\forall\ \boldsymbol{\beta}\in\mathbb{B}_1$, $u(\cdot|\boldsymbol{\beta})\in C_1^1(T)$, we can find $l,\ l'$ and a such that $\boldsymbol{\beta}\in\mathbb{C}(\boldsymbol{\beta}_{ll'}^\circ)$ and $u_a^\circ(\cdot)$ satisfies

$$n^{-1}\sum_{i=1}^n\left[u(\boldsymbol{x}_i^T\boldsymbol{\beta}_1|\boldsymbol{\beta}_1)-u_a^\circ(\boldsymbol{x}_i^T\boldsymbol{\beta}_1)\right]^2\leqslant\delta_n^2,\ \forall\boldsymbol{\beta}_1\in\mathbb{B}_1. \tag{S41}$$

On the event $\mathcal{G}_n\cap\mathcal{J}_n\cap\mathcal{K}_n$, we have

$$\begin{aligned}
&n^{-1}\left|\sum_{i=1}^n\boldsymbol{\theta}_j^T\boldsymbol{\nu}_{21}(Z_i,\boldsymbol{\beta},u)-\sum_{i=1}^n\boldsymbol{\theta}_j^T\boldsymbol{\nu}_{21}(Z_i,\boldsymbol{\beta}_{ll'}^\circ,u_a^\circ)\right|\\
\leqslant&n^{-1}\left|\sum_{i=1}^n\left[G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-G(\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^\circ|\boldsymbol{\beta}_{ll'}^\circ)\right]G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})[\boldsymbol{x}_{i,-1}-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})]^T\boldsymbol{\theta}_j\right|\\
&+n^{-1}\left|\sum_{i=1}^n\left[u(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-u_a^\circ(\boldsymbol{x}_i^T\boldsymbol{\beta})\right]G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})[\boldsymbol{x}_{i,-1}-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})]^T\boldsymbol{\theta}_j\right|\\
&+n^{-1}\left|\sum_{i=1}^n\left[u_a^\circ(\boldsymbol{x}_i^T\boldsymbol{\beta})-u_a^\circ(\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^\circ)\right]G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})[\boldsymbol{x}_{i,-1}-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})]^T\boldsymbol{\theta}_j\right|\\
&+n^{-1}\left|\sum_{i=1}^n\left[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^\circ|\boldsymbol{\beta}_{ll'}^\circ)-u(\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^\circ|\boldsymbol{\beta}_{ll'}^\circ)\right]\left[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^\circ|\boldsymbol{\beta}_{ll'}^\circ)\right][\boldsymbol{x}_{i,-1}-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})]^T\boldsymbol{\theta}_j\right|\\
&+n^{-1}\left|\sum_{i=1}^n\left[u(\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^\circ|\boldsymbol{\beta}_{ll'}^\circ)-u_a^\circ(\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^\circ)\right]\left[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^\circ|\boldsymbol{\beta}_{ll'}^\circ)\right][\boldsymbol{x}_{i,-1}-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})]^T\boldsymbol{\theta}_j\right|\\
&+n^{-1}\left|\sum_{i=1}^n\left[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^\circ|\boldsymbol{\beta}_{ll'}^\circ)-u(\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^\circ|\boldsymbol{\beta}_{ll'}^\circ)\right]G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^\circ|\boldsymbol{\beta}_{ll'}^\circ)[\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^\circ)-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})]^T\boldsymbol{\theta}_j\right|\\
&+n^{-1}\left|\sum_{i=1}^n\left[u(\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^\circ|\boldsymbol{\beta}_{ll'}^\circ)-u_a^\circ(\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^\circ)\right]G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^\circ|\boldsymbol{\beta}_{ll'}^\circ)[\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^\circ)-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})]^T\boldsymbol{\theta}_j\right|\\
=&\sum_{j=1}^7R_{nj},
\end{aligned}$$

where the definition of R_{nj}, $j = 1, \cdots, 7$, is clear from the context. (S4) in Lemma A4 and the event $\mathcal{G}_n$ implies that $R_{n1} \leqslant c_1\sqrt{||\boldsymbol{\beta} - \boldsymbol{\beta}^{\circ}_{ll'}||_2 s\log(p \vee n)} \leqslant d_0\delta_n$, with probability at least $1 - \exp[-c_2 \log(p \vee n)]$, for some positive constants d_0, c_1, c_2, and all n sufficiently large. The Cauchy-Schwartz inequality implies that

$$R_{n2} \leqslant \sqrt{n^{-1}\sum_{i=1}^{n}\left[u(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - u_a^{\circ}(\boldsymbol{x}_i^T\boldsymbol{\beta}]\right]^2}\sqrt{n^{-1}\sum_{i=1}^{n}\left\{G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_{i,-1}^T\boldsymbol{\beta})]^T\boldsymbol{\theta}_j\right\}^2} \leqslant d_0\delta_n,$$

for some positive constant d_0, which applies the event $\mathcal{G}_n$, Assumption (A5)-(a) and (S41). Since $u_a^{\circ}(\cdot) \in C_1^1(T)$, we have

$$R_{n3} \leqslant d_1\sqrt{n^{-1}\sum_{i=1}^{n}\left[\boldsymbol{x}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}^{\circ}_{ll'})\right]^2}\sqrt{n^{-1}\sum_{i=1}^{n}\left\{G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_{i,-1}^T\boldsymbol{\beta})]^T\boldsymbol{\theta}_j\right\}^2} \leqslant d_0\delta_n,$$

for some positive constants d_0, d_1, which applies the event $\mathcal{G}_n \cap \mathcal{K}_n$, and Assumption (A5)-(a). Similarly, we can derive that $R_{n4} \leqslant d_0h^2\delta_n$ and $R_{n5} \leqslant d_0\delta_n^2$, with probability at least $1 - \exp[-c_2\log(p \vee n)]$, for some positive constants d_0, c_2, and all n sufficiently large, which applies (S5) in Lemma A4 and the event $\mathcal{G}_n$. In addition, we have $R_{n6} \leqslant d_0h^2\delta_n$ and $R_{n7} \leqslant d_0\delta_n^2$, which both apply Assumption (A2)-(c) and Assumption (A5)-(a).

Hence, the δ_n-covering number of the class of functions $V_{2j} = \{\boldsymbol{\theta}_j^T\boldsymbol{\nu}_{21}(Z_i, \boldsymbol{\beta}, u) : \boldsymbol{\beta} \in \mathbb{B}_1, u \in \mathbb{U}\}$ is bounded by

$$N(\delta_n, V_{2j}, L_1(\mathbb{P}_n)) \leqslant c\binom{p}{ks}\left[1 + \frac{2c_0s^{3/2}h^2\log(p \vee n)}{\delta_n^2}\right]^{ks}\exp\left[C||\boldsymbol{\beta}_0||_2\delta_n^{-1}\sqrt{\log(p \vee n)}\right], \ \forall j. \tag{S42}$$

Let $a_1, \cdots, a_n$ be a Rademacher sequence that is independent of the data. The symmetrization theorem (Theorem 14.3 in Bühlmann and van de Geer [2011]) implies that

$$\begin{aligned}E_{sup,2} =&\mathrm{E}\left[\sup_{\boldsymbol{\beta}\in\mathbb{B}_1,u\in\mathbb{U}}\left|n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\theta}_j^T\boldsymbol{\nu}_{21}(Z_i,\boldsymbol{\beta},u)\right|\,\middle|\mathcal{G}_n\cap\mathcal{J}_n\cap\mathcal{K}_n\right]\\ \leqslant&2\mathrm{E}_{\{\boldsymbol{x}_i\}_{i=1}^n}\left\{\mathrm{E}\left[\sup_{\boldsymbol{\beta}\in\mathbb{B}_1,u\in\mathbb{U}}\left|n^{-1/2}\sum_{i=1}^{n}a_i\boldsymbol{\theta}_j^T\boldsymbol{\nu}_{21}(Z_i,\boldsymbol{\beta},u)\right|\,\middle|\{\boldsymbol{x}_i\}_{i=1}^n,\mathcal{G}_n\cap\mathcal{J}_n\cap\mathcal{K}_n\right]\right\}.\end{aligned}$$

Note that on the event $\mathcal{G}_n\cap\mathcal{J}_n$, we have that $\sup_{\boldsymbol{\beta}\in\mathbb{B}_1,u\in\mathbb{U}}\sqrt{n^{-1}\sum_{i=1}^n[a_i\boldsymbol{\theta}_j^T\boldsymbol{\nu}_{21}(Z_i,\boldsymbol{\beta},u)]^2}\leqslant ch^2\triangleq R_n$. Lemma 14.18 in van der Vaart and Wellner [1996] implies that

$$\begin{aligned}E_{sup,2} \leqslant&2\mathrm{E}_{\{\boldsymbol{x}_i\}_{i=1}^n}\left\{\mathrm{E}\left[\sup_{\boldsymbol{\beta}\in\mathbb{B}_1,u\in\mathbb{U}}\left|n^{-1/2}\sum_{i=1}^{n}a_i\boldsymbol{\theta}_j^T\boldsymbol{\nu}_{21}(Z_i,\boldsymbol{\beta},u)\right|\,\middle|\{\boldsymbol{x}_i\}_{i=1}^n,\mathcal{G}_n\cap\mathcal{J}_n\cap\mathcal{K}_n\right]\right\}\\ \leqslant&ch^2\left\{4+6\sum_{l=1}^{L}2^{-l}\sqrt{\log\left[N(2^{-l}h^2,V_j,L_1(\mathbb{P}_n))+1\right]}\right\}\\ \leqslant&c_0h[\log(p\vee n)]^{1/4},\end{aligned}$$

with probability at least $1-\exp[-c_1\log(p\vee n)]$, for some positive constants c, c_0, c_1, and all n sufficiently large, where $L=\min\{l:l\geqslant 1,\ 2^{-l}\leqslant 4/\sqrt{n}\}$, and the last inequality applies (S42). The analysis is similar as that in the proof of Lemma A8.

It follows from (S40) that

$$\begin{aligned}&P\left(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1,u\in\mathbb{U}}\left|n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\theta}_j^T\boldsymbol{\nu}_{21}(Z_i,\boldsymbol{\beta},u)\right|\geqslant ch[\log(p\vee n)]^{1/4}+th^2\sqrt{n}\,\middle|\mathcal{G}_n\cap\mathcal{J}_n\cap\mathcal{K}_n\right)\\ \leqslant&\exp\left(-\frac{nt^2}{8}\right).\end{aligned}$$

Take $t=4\sqrt{n^{-1}\log(p\vee n)}$. Note that the assumptions of Theorem 1 imply $h[\log(p\vee n)]^{1/4}\leqslant d_1(nh^9)^{1/4}\leqslant d_1$, for some positive constant d_1. Hence there exist some positive constants c_0, c_1, such that for all n sufficiently large,

$$P\left(\max_{2\leqslant j\leqslant p}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1,u\in\mathbb{U}}\left|n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\theta}_j^T\boldsymbol{\nu}_{21}(Z_i,\boldsymbol{\beta},u)\right|\geqslant c_0h[\log(p\vee n)]^{1/4}\right)\leqslant\exp[-c_1\log(p\vee n)].$$

Therefore, combining all the above results, we prove (S11).

$\square$

Proof of Lemma A10. We will prove (S13) below. The proof of (S14) is similar.

Theorem 1 implies that $P\left(\widehat{\boldsymbol{\beta}} \in \mathbb{B}_1\right) \geqslant 1 - \exp(-c \log p)$, for some positive constant c, and all n sufficiently large. Lemma A11 implies that

$$P\left(\max_{1\leqslant i\leqslant n} \sup_{\boldsymbol{v}\in\mathbb{K}(p,2ks+\widetilde{s})} \left|\widehat{\mathrm{E}}(\boldsymbol{x}_i^T\boldsymbol{v}|\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}) - \mathrm{E}(\boldsymbol{x}_i^T\boldsymbol{v}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\right| \geqslant c_0 sh^2\sqrt{\log(p\vee n)}\right) \leqslant \exp(-c_1\log p),$$

for some positive constants c_0, c_1 and all n sufficiently large. Define

$$\begin{aligned}\mathbb{E} = \Big\{&\boldsymbol{E}(\cdot|\boldsymbol{\beta}) : \boldsymbol{\beta}\in\mathbb{B}_1;\ \text{for any unit vector } \boldsymbol{v}\in\mathbb{R}^p, \boldsymbol{E}(\cdot|\boldsymbol{\beta})^T\boldsymbol{v}\in C_1^1(T),\ \forall\ \boldsymbol{\beta},\\ &\text{and} \max_{1\leqslant i\leqslant n} \sup_{\substack{\boldsymbol{\beta}\in\mathbb{B}_1\\ \boldsymbol{v}\in\mathbb{K}(p,2ks+\widetilde{s})}} \left|[\boldsymbol{E}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - \mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^T\boldsymbol{v}\right| \leqslant c_1 sh^2\sqrt{\log(p\vee n)}\Big\},\end{aligned}$$

for some positive constant c_1, where $T = \left\{t\in\mathbb{R} : |t| \leqslant 2||\boldsymbol{\beta}_0||_2\sigma_x\sqrt{\log(p\vee n)}\right\}$, and $C_1^1(T)$ is the set of all continuous and Lipschitz functions $f : T \to \mathbb{R}$. To prove (S13), it is sufficient to prove that there exist some positive constants c_0, c_1, such that for all n sufficiently large,

$$P\left(\max_{2\leqslant j\leqslant p} \sup_{\boldsymbol{\beta}\in\mathbb{B}_1,\boldsymbol{E}\in\mathbb{E}} \left|n^{-1/2}\sum_{i=1}^n \boldsymbol{\theta}_j^T\boldsymbol{\xi}(Z_i,\boldsymbol{\beta},\boldsymbol{E})\right| \geqslant c_0 h\sqrt{s\log(p\vee n)}\right) \leqslant \exp[-c_1\log(p\vee n)],$$

where $\boldsymbol{\xi}(Z_i,\boldsymbol{\beta},\boldsymbol{E}) = \widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)[\boldsymbol{E}_{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]$, $\boldsymbol{E}(\boldsymbol{x}^T\boldsymbol{\beta}|\boldsymbol{\beta})$ depends on $\boldsymbol{x}$ only through $\boldsymbol{x}^T\boldsymbol{\beta}$, and $\boldsymbol{E}_{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})$ denotes the $(p-1)-$subvector of $\boldsymbol{E}(\boldsymbol{x}^T\boldsymbol{\beta}|\boldsymbol{\beta})$ that excludes the 1^{st} entry.

Note that $\max_{1\leqslant i\leqslant n}\left|G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right| \leqslant b$ according to assumption (A1)-(b). Using the same technique as that in Lemma A3, the event $\mathcal{Q}_n = \left\{n^{-1}\sum_{i=1}^n\left[\widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\right]^2 \leqslant 5b^2(\sigma_\epsilon^2 + M^2)\right\}$ holds with probability at least $1-\exp(-cn)$, for some universal constant $c > 0$. Then there exist some positive constant c, such that for all n sufficiently large,

$$\begin{aligned} &P\left(\max_{2\leqslant j\leqslant p}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1,\boldsymbol{E}\in\mathbb{E}}\left|n^{-1/2}\sum_{i=1}^n\boldsymbol{\theta}_j^T\boldsymbol{\xi}(Z_i,\boldsymbol{\beta},\boldsymbol{E})\right|>t\right)\\ &\leqslant\sum_{j=2}^p P\left(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1,\boldsymbol{E}\in\mathbb{E}}\left|n^{-1/2}\sum_{i=1}^n\boldsymbol{\theta}_j^T\boldsymbol{\xi}(Z_i,\boldsymbol{\beta},\boldsymbol{E})\right|>t\;\middle|\mathcal{Q}_n\cap\mathcal{J}_n\cap\mathcal{K}_n\right)+\exp[-c\log(p\vee n)]. \end{aligned}$$

Note that $\boldsymbol{\theta}_j^T\boldsymbol{\xi}(Z_i,\boldsymbol{\beta},\boldsymbol{E})=2(2A_i-1)\left[\epsilon_i+g(\boldsymbol{x}_i)\right]G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\boldsymbol{\theta}_j^T\left[\boldsymbol{E}_{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\right]$, where $(2A_i-1)$ is a Rademacher sequence, and independent of $(\boldsymbol{x}_i,\epsilon_i)$. Hence given $\{(\boldsymbol{x}_i,\epsilon_i)\}_{i=1}^n$, on the event $\mathcal{Q}_n$, we have

$$\sup_{\boldsymbol{\beta}\in\mathbb{B}_1,\boldsymbol{E}\in\mathbb{E}}n^{-1}\sum_{i=1}^n\left|\boldsymbol{\theta}_j^T\boldsymbol{\xi}(Z_i,\boldsymbol{\beta},\boldsymbol{E})\right|^2\leqslant Cs^2h^4\log(p\vee n),$$

for some positive constant C, and any j. Hence by Massart's concentration inequality (e.g., Theorem 14.2, Bühlmann and van de Geer [2011]) on $\mathcal{Q}_n$, given $\{(\boldsymbol{x}_i,\epsilon_i)\}_{i=1}^n$, $\forall\ t>0$,

$$\begin{aligned} &P\Bigg(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1,\boldsymbol{E}\in\mathbb{E}}\left|n^{-1/2}\sum_{i=1}^n\boldsymbol{\theta}_j^T\boldsymbol{\xi}(Z_i,\boldsymbol{\beta},\boldsymbol{E})\right|\geqslant\mathrm{E}_{sup}\big[\boldsymbol{\theta}_j^T\boldsymbol{\xi}(Z_i,\boldsymbol{\beta},\boldsymbol{E})\big]+tsh^2\sqrt{n\log(p\vee n)}\\ &\qquad\qquad\qquad\qquad\qquad\qquad\Bigg|\{(\boldsymbol{x}_i,\epsilon_i)\}_{i=1}^n,\mathcal{Q}_n\cap\mathcal{J}_n\cap\mathcal{K}_n\Bigg)\\ &\leqslant\exp\left(-\frac{nt^2}{8}\right), \end{aligned}\tag{S43}$$

where $\mathrm{E}_{sup}\big[\boldsymbol{\theta}_j^T\boldsymbol{\xi}(Z_i,\boldsymbol{\beta},\boldsymbol{E})\big]=\mathrm{E}\left[\sup_{\boldsymbol{\beta}\in\mathbb{B}_1,\boldsymbol{E}\in\mathbb{E}}\left|n^{-1/2}\sum_{i=1}^n\boldsymbol{\theta}_j^T\boldsymbol{\xi}(Z_i,\boldsymbol{\beta},\boldsymbol{E})\right|\;\middle|\{(\boldsymbol{x}_i,\epsilon_i)\}_{i=1}^n,\mathcal{Q}_n\cap\mathcal{J}_n\cap\mathcal{K}_n\right]$ It remains to find an upper bound for $\mathrm{E}_{sup}\big[\boldsymbol{\theta}_j^T\boldsymbol{\xi}(Z_i,\boldsymbol{\beta},\boldsymbol{E})\big]$.

Use the δ_n-cover for $\mathbb{B}_1$ and $C_1^1(T)$ constructed in the proof of Lemma A8. Hence $\forall\ \boldsymbol{\beta}\in\mathbb{B}_1$, $\big[\boldsymbol{\theta}_j^T\boldsymbol{E}_{-1}(\cdot|\boldsymbol{\beta})\big]\in C_1^1(T)$ for any $j=2,\cdots,p$, we can find l, l' and a such that $\boldsymbol{\beta}\in\mathbb{C}(\boldsymbol{\beta}_{ll'}^\circ)$ and $\boldsymbol{E}_a^\circ(\cdot)$ satisfies

$$n^{-1}\sum_{i=1}^n\left\{[\boldsymbol{E}_{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}_1|\boldsymbol{\beta})-\boldsymbol{E}_{a,-1}^\circ(\boldsymbol{x}_i^T\boldsymbol{\beta}_1)]^T\boldsymbol{\theta}_j\right\}^2\leqslant\delta_n^2,\quad\forall\boldsymbol{\beta}_1\in\mathbb{B}_1.$$

On the event $\mathcal{Q}_n \cap \mathcal{J}_n \cap \mathcal{K}_n$, there exists some positive constant c, such that

$$
\begin{aligned}
& n^{-1}\left|\sum_{i=1}^{n} \boldsymbol{\theta}_j^T \boldsymbol{\xi}(Z_i, \boldsymbol{\beta}, \boldsymbol{E}) - \sum_{i=1}^{n} \boldsymbol{\theta}_j^T \boldsymbol{\xi}(Z_i, \boldsymbol{\beta}_{ll'}^{\circ}, \boldsymbol{E}_a^{\circ})\right| \\
\leqslant & n^{-1}\left|\sum_{i=1}^{n} \boldsymbol{\theta}_j^T \left[\boldsymbol{E}_{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - \boldsymbol{E}_{a,-1}^{\circ}(\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^{\circ})\right] \widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right| \\
\leqslant & \sqrt{n^{-1}\sum_{i=1}^{n}\left\{\boldsymbol{\theta}_j^T\left[\boldsymbol{E}_{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - \boldsymbol{E}_{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^{\circ}|\boldsymbol{\beta})\right]\right\}^2 + \left\{\boldsymbol{\theta}_j^T\left[\boldsymbol{E}_{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^{\circ}|\boldsymbol{\beta}) - \boldsymbol{E}_{a,-1}^{\circ}(\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^{\circ})\right]\right\}^2} \\
& * \sqrt{n^{-1}\sum_{i=1}^{n}\left[\widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right]^2} \\
\leqslant & c\delta_n.
\end{aligned}
$$

Hence, the δ_n-covering number of the class of functions $\Phi_j = \{\boldsymbol{\theta}_j^T\boldsymbol{\xi}(Z_i, \boldsymbol{\beta}, \boldsymbol{E}) : \boldsymbol{\beta} \in \mathbb{B}_1, \boldsymbol{E} \in \mathbb{E}\}$ satisfies

$$
N(\delta_n, \Phi_j, L_1(\mathbb{P}_n)) \leqslant c\binom{p}{ks}\left(1 + \frac{2c_0\sqrt{s}h^2}{\delta_n}\right)^{ks} \exp\left[C||\boldsymbol{\beta}_0||_2\delta_n^{-1}\sqrt{\log(p \vee n)}\right], \quad \forall j. \qquad \text{(S44)}
$$

Recall $\boldsymbol{\theta}_j^T\boldsymbol{\xi}(Z_i, \boldsymbol{\beta}, \boldsymbol{E}) = 2(2A_i - 1)\left[\epsilon_i + g(\boldsymbol{x}_i)\right] G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\boldsymbol{\theta}_j^T\left[\boldsymbol{E}_{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\right]$, where $(2A_i - 1)$ is a Rademacher sequence, and independent of $(\boldsymbol{x}_i, \epsilon_i)$. Note that on the event $\mathcal{Q}_n \cap \mathcal{J}_n \cap \mathcal{K}_n$, we have that $\sup_{\boldsymbol{\beta}\in\mathbb{B}_1, \boldsymbol{E}\in\mathbb{E}} \sqrt{n^{-1}\sum_{i=1}^{n}\left[\boldsymbol{\theta}_j^T\boldsymbol{\xi}(Z_i, \boldsymbol{\beta}, \boldsymbol{E})^2\right]} = c_0sh^2\sqrt{\log(p \vee n)} \triangleq R_n$. Therefore, Lemma 14.18 in van der Vaart and Wellner [1996] implies that

$$
\begin{aligned}
& \mathrm{E}_{sup}\left[\boldsymbol{\theta}_j^T\boldsymbol{\xi}(Z_i, \boldsymbol{\beta}, \boldsymbol{E})\right] \\
= & \mathrm{E}\left[\sup_{\boldsymbol{\beta}\in\mathbb{B}_1, \boldsymbol{E}\in\mathbb{E}}\left|n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\theta}_j^T\boldsymbol{\xi}(Z_i, \boldsymbol{\beta}, \boldsymbol{E})\right| \,\middle|\, \{(\boldsymbol{x}_i, \epsilon_i)\}_{i=1}^n, \mathcal{Q}_n \cap \mathcal{J}_n \cap \mathcal{K}_n\right] \\
\leqslant & Csh^2\sqrt{\log(p \vee n)}\left\{4 + 6\sum_{l=1}^{L} 2^{-l}\sqrt{\log\left[N(2^{-l}sh^2\sqrt{\log(p \vee n)}, \Phi_j, L_1(\mathbb{P}_n)) + 1\right]}\right\} \\
\leqslant & c_0h\sqrt{s\log(p \vee n)},
\end{aligned}
$$

for some positive constants C, c_0, and all n sufficiently large, where $L = \min\{l : l \geqslant 1,\ 2^{-l} \leqslant$

$4/\sqrt{n}\}$, and the last inequality applies (S44). It follows from (S43) that

$$P\left(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1,\boldsymbol{E}\in\mathbb{E}}\left|n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\theta}_j^T\boldsymbol{\xi}(Z_i,\boldsymbol{\beta},\boldsymbol{E})\right|\geqslant ch\sqrt{s\log(p\vee n)}+tsh^2\sqrt{n\log(p\vee n)}\right.$$
$$\left.\Big|\{(\boldsymbol{x}_i,\epsilon_i)\}_{i=1}^n,\mathcal{Q}_n\cap\mathcal{J}_n\cap\mathcal{K}_n\right)\leqslant\exp\left(-\frac{nt^2}{8}\right).$$

Take $t=4\sqrt{n^{-1}\log(p\vee n)}$. Note that the assumptions of Theorem 1 imply $h\sqrt{s\log(p\vee n)}\leqslant c_1\sqrt{nh^7}\leqslant c_1$, for some positive constant c_1. Hence we have

$$\begin{aligned}&P\left(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1,\boldsymbol{E}\in\mathbb{E}}\left|n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\theta}_j^T\boldsymbol{\xi}(Z_i,\boldsymbol{\beta},\boldsymbol{E})\right|\geqslant ch\sqrt{s\log(p\vee n)}\ \Big|\mathcal{Q}_n\cap\mathcal{J}_n\cap\mathcal{K}_n\right)\\
=&\mathrm{E}\left\{P\left(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1,\boldsymbol{E}\in\mathbb{E}}\left|n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\theta}_j^T\boldsymbol{\xi}(Z_i,\boldsymbol{\beta},\boldsymbol{E})\right|\geqslant ch\sqrt{s\log(p\vee n)}\ \Big|\{(\boldsymbol{x}_i,\epsilon_i)\}_{i=1}^n,\mathcal{Q}_n\cap\mathcal{J}_n\cap\mathcal{K}_n\right)\right\}\\
\leqslant&\exp\left[-2\log(p\vee n)\right].\end{aligned}$$

Therefore, there exist some positive constants c_0, c_1, such that for all n sufficiently large,

$$P\left(\max_{2\leqslant j\leqslant p}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1,\boldsymbol{E}\in\mathbb{E}}\left|n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\theta}_j^T\boldsymbol{\xi}(Z_i,\boldsymbol{\beta},\boldsymbol{E})\right|\geqslant c_0h\sqrt{s\log(p\vee n)}\right)\leqslant\exp[-c_1\log(p\vee n)].$$

□

Proof of Lemma A11. We will prove (S15) and (S16) below. The proof of (S17) is similar.

$$\begin{aligned}&\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1n}\left|\widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right|\\
\leqslant&\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|\widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\right|\\
&+\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right|.\end{aligned}$$

Lemma A4 implies that

$$\begin{aligned}
&\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right|\\
\leqslant&\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|f_0'(\boldsymbol{x}_i^T\boldsymbol{\beta})[\boldsymbol{x}_{i,-1}-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})]^T\boldsymbol{\gamma}\right|+\frac{1}{2}\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|f_0''(\boldsymbol{x}_i^T\boldsymbol{\beta}_1)(\boldsymbol{x}_{i,-1}^T\boldsymbol{\gamma})^2\right|\\
&+\frac{1}{2}\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|\mathrm{E}\big[f_0''(\boldsymbol{x}_i^T\boldsymbol{\beta}_2)(\boldsymbol{x}_{i,-1}^T\boldsymbol{\gamma})^2|\boldsymbol{x}_i^T\boldsymbol{\beta}\big]\right|\\
\leqslant& C\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|[\boldsymbol{x}_{i,-1}-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})]^T\boldsymbol{\gamma}\right|+C\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}(\boldsymbol{x}_{i,-1}^T\boldsymbol{\gamma})^2\\
&+C\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|\mathrm{E}\big[(\boldsymbol{x}_{i,-1}^T\boldsymbol{\gamma})^2|\boldsymbol{x}_i^T\boldsymbol{\beta}\big]\right|\\
\leqslant& C\max_{1\leqslant i\leqslant n}\left|\left|\boldsymbol{x}_i-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})\right|\right|_\infty\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}||\boldsymbol{\gamma}||_1+C\max_{1\leqslant i\leqslant n}||\boldsymbol{x}_i||_\infty^2\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}||\boldsymbol{\gamma}||_1^2+C||\boldsymbol{\gamma}||_2^2\log(p\vee n),
\end{aligned}$$

where $\boldsymbol{\gamma}=\boldsymbol{\beta}_{-1}-\boldsymbol{\beta}_{0,-1}$, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$. The last part of the last inequality comes from Assumption (A2)-(b).

For a sub-Gaussian random vector $\boldsymbol{x}\in\mathbb{R}^p$, its property implies that

$$P\left(\max_{1\leqslant i\leqslant n}||\boldsymbol{x}_i||_\infty\geqslant c_1\sqrt{\log(p\vee n)}\right)\leqslant\exp\left[-c_2\log(p\vee n)\right],$$

for positive constants c_1 and c_2. Similar bounds also work for $\max_{1\leqslant i\leqslant n}\left|\left|\boldsymbol{x}_i-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})\right|\right|_\infty$.The assumptions of Theorem 1 imply that $||\boldsymbol{\gamma}||_2^2\log(p\vee n)\leqslant d_0sh^4\log(p\vee n)=d_0sh^2\sqrt{\log(p\vee n)}*\sqrt{h^4\log(p\vee n)}\leqslant d_1sh^2\sqrt{\log(p\vee n)}\sqrt{nh^9}\leqslant d_1sh^2\sqrt{\log(p\vee n)}$, for some positive constants d_0, d_1. Hence we have

$$\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right|\leqslant Csh^2\sqrt{\log(p\vee n)},$$

for positive constant C, with probability at least $1-\exp[-c_2\log(p\vee n)]$. Hence we can apply

Lemma A5 and derive that

$$P\left(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|\widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right|\geqslant c_0sh^2\sqrt{\log(p\vee n)}\right)\leqslant\exp[-c_1\log(p\vee n)].$$

Hence we can conclude (S15).

Similarly, it's sufficient to bound $\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right|$ to prove (S16). Lemma A4 implies that there exist $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$, such that

$$\begin{aligned}
&\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right|\\
\leqslant&\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|f_0''(\boldsymbol{x}_i^T\boldsymbol{\beta})[\boldsymbol{x}_{i,-1}-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})]^T\boldsymbol{\gamma}\right|+\frac{1}{2}\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|f_0'''(\boldsymbol{x}_i^T\boldsymbol{\beta}_1)(\boldsymbol{x}_{i,-1}^T\boldsymbol{\gamma})^2\right|\\
&+\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|f_0'(\boldsymbol{x}_i^T\boldsymbol{\beta})\mathrm{E}^{(1)}(\boldsymbol{x}_{i,-1}^T\boldsymbol{\gamma}|\boldsymbol{x}_i^T\boldsymbol{\beta})\right|+\frac{1}{2}\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|\mathrm{E}^{(1)}\left[f_0''(\boldsymbol{x}_i^T\boldsymbol{\beta}_2)(\boldsymbol{x}_{i,-1}^T\boldsymbol{\gamma})^2|\boldsymbol{x}_i^T\boldsymbol{\beta}\right]\right|\\
\leqslant&C\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|[\boldsymbol{x}_{i,-1}-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})]^T\boldsymbol{\gamma}\right|+C\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}(\boldsymbol{x}_{i,-1}^T\boldsymbol{\gamma})^2\\
&+C\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|\mathrm{E}^{(1)}(\boldsymbol{x}_{i,-1}^T\boldsymbol{\gamma}|\boldsymbol{x}_i^T\boldsymbol{\beta})\right|+C\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|\mathrm{E}^{(1)}\left[(\boldsymbol{x}_{i,-1}^T\boldsymbol{\gamma})^2|\boldsymbol{x}_i^T\boldsymbol{\beta}\right]\right|\\
\leqslant&C\max_{1\leqslant i\leqslant n}\left|\left|\boldsymbol{x}_i-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})\right|\right|_\infty\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}||\boldsymbol{\gamma}||_1+C\max_{1\leqslant i\leqslant n}||\boldsymbol{x}_i||_\infty\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}||\boldsymbol{\gamma}||_1^2+C||\boldsymbol{\gamma}||_2+C||\boldsymbol{\gamma}||_2^2\sqrt{\log(p\vee n)}\\
\leqslant&c_0sh^2\sqrt{\log(p\vee n)},
\end{aligned}$$

with probability at least $1-\exp[-c_1\log(p\vee n)]$, for some positive constants C, c_0, c_1, and all n sufficiently large. The last two parts of the second last inequality come from Assumption (A2)-(b). Note that $sh^2\sqrt{\log(p\vee n)}\leqslant d_0nh^7\leqslant d_0h$ for some positive constant d_0. We thus have

$$P\left(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left|G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right|\geqslant c_0h\right)\leqslant\exp[-c_1\log(p\vee n)],$$

for some positive constants c_0, c_1, and all n sufficiently large. Combining this result with Lemma A6, we can conclude (S16). $\square$

Proof of Lemma A12. We will prove (S18) below. The proof of (S19) is similar.

$$\begin{aligned}
&\sup_{\substack{\boldsymbol{\beta}\in\mathbb{B}_1\\ \boldsymbol{v}\in\mathbb{K}(p,2ks+\widetilde{s})}} \Big|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{v}^T[\widehat{\mathrm{E}}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)][\widehat{\mathrm{E}}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^T\boldsymbol{v}\Big| \\
\leqslant\; &\sup_{\substack{\boldsymbol{\beta}\in\mathbb{B}_1\\ \boldsymbol{v}\in\mathbb{K}(p,2ks+\widetilde{s})}} \Big|\frac{2}{n}\sum_{i=1}^{n}\boldsymbol{v}^T[\widehat{\mathrm{E}}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)][\widehat{\mathrm{E}}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^T\boldsymbol{v}\Big| \\
&+\sup_{\substack{\boldsymbol{\beta}\in\mathbb{B}_1\\ \boldsymbol{v}\in\mathbb{K}(p,2ks+\widetilde{s})}} \Big|\frac{2}{n}\sum_{i=1}^{n}\boldsymbol{v}^T[\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)][\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^T\boldsymbol{v}\Big|.
\end{aligned}$$

Lemma A7 implies that $\sup_{\substack{\boldsymbol{\beta}\in\mathbb{B}\\ \boldsymbol{v}\in\mathbb{K}(p,2ks+\widetilde{s})}} \frac{1}{n}\sum_{i=1}^{n}\big|[\widehat{\mathrm{E}}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})]^T\boldsymbol{v}\big|^2 \leqslant c_0h^4$ holds with probability at least $1-\exp[-c_1\log(p\vee n)]$, for some positive constants c_0, c_1, and all n sufficiently large.

Assumption (A2) implies that there exists positive constant C such that

$$\begin{aligned}
&\sup_{\substack{\boldsymbol{\beta}\in\mathbb{B}_1\\ \boldsymbol{v}\in\mathbb{K}(p,2ks+\widetilde{s})}} \frac{1}{n}\sum_{i=1}^{n}\big|[\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^T\boldsymbol{v}\big|^2 \\
\leqslant &C\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\frac{1}{n}\sum_{i=1}^{n}\big|\boldsymbol{x}_i^T\boldsymbol{\beta}-\boldsymbol{x}_i^T\boldsymbol{\beta}_0|^2 + C\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\frac{1}{n}\sum_{i=1}^{n}(|\boldsymbol{x}_i^T\boldsymbol{\beta}|^2+|\boldsymbol{x}_i^T\boldsymbol{\beta}_0|^2)||\boldsymbol{\beta}-\boldsymbol{\beta}_0||_2^2.
\end{aligned}$$

By Lemma B2, we have that $\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\frac{1}{n}\sum_{i=1}^{n}\big|\boldsymbol{x}_i^T\boldsymbol{\beta}|^2 * ||\boldsymbol{\beta}-\boldsymbol{\beta}_0||_2^2 \leqslant c_0sh^4$ with probability at least $1-\exp[-c_1\log(p\vee n)]$, for some positive constants c_0, c_1, and all n sufficiently large. Similarly, we can derive that $\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\frac{1}{n}\sum_{i=1}^{n}\big|\boldsymbol{x}_i^T\boldsymbol{\beta}-\boldsymbol{x}_i^T\boldsymbol{\beta}_0|^2 \leqslant c_1sh^4$with probability at least $1-\exp[-c_1\log(p\vee n)]$, for some positive constants c_0, c_1, and all n sufficiently large. Combining these two results, we can derive that

$$P\Big(\sup_{\substack{\boldsymbol{\beta}\in\mathbb{B}\\ \boldsymbol{v}\in\mathbb{K}(p,2ks+\widetilde{s})}} \frac{1}{n}\sum_{i=1}^{n}\big|[\widehat{\mathrm{E}}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^T\boldsymbol{v}\big|^2 \geqslant c_0sh^4\Big) \leqslant \exp(-c_1\log p),$$

for some positive constants c_0, c_1, and all n sufficiently large. Hence we can conclude (S18).

□

Proof of Lemma A13. First, we define $\boldsymbol{\phi}_{0j}^* \triangleq \Big(-(\boldsymbol{d}_{0j})_{1:(j-2)}^T, 1, -(\boldsymbol{d}_{0j})_{(j-1):(p-2)}^T \Big)^T$. By the definitions of $\boldsymbol{d}_{0j}$ and τ_{0j}^2, we know that

$$\boldsymbol{\Omega}_{-(j-1)}\boldsymbol{\phi}_{0j}^* = \boldsymbol{\Omega}_{-(j-1),(j-1)} - \boldsymbol{\Omega}_{-(j-1),-(j-1)}\boldsymbol{d}_{0j} = \boldsymbol{0}_{p-2},$$
$$\boldsymbol{\Omega}_{j-1}^T\boldsymbol{\phi}_{0j}^* = \Omega_{(j-1),(j-1)} - \boldsymbol{\Omega}_{(j-1),-(j-1)}\boldsymbol{d}_{0j} = \tau_{0j}^2,$$

where $\boldsymbol{\Omega}_{j-1} \in \mathbb{R}^{p-1}$ is the $(j-1)^{th}$ column of $\boldsymbol{\Omega}$, and $\boldsymbol{\Omega}_{-(j-1)} \in \mathbb{R}^{(p-2)\times(p-1)}$ is the submatrix of $\boldsymbol{\Omega}$ with its $(j-1)^{th}$ row removed. Given these two facts, we can derive that

$$\boldsymbol{\Omega}\boldsymbol{\phi}_{0j}^* = \tau_{0j}^2\boldsymbol{e}_{j-1} = \boldsymbol{\Omega}\boldsymbol{\theta}_j\tau_{0j}^2,$$

since $\boldsymbol{\Omega}\boldsymbol{\Theta} = \boldsymbol{I}_{p-1}$, where $\boldsymbol{e}_{j-1}$ is the $(j-1)^{th}$ column of $\boldsymbol{I}_{p-1}$. Assumption (A2)-(a) indicates that $\lambda_{min}(\boldsymbol{\Omega}) \geqslant \xi_2 > 0$, then we have $\boldsymbol{\phi}_{0j}^* = \Big(-(\boldsymbol{d}_{0j})_{1:(j-2)}^T, 1, -(\boldsymbol{d}_{0j})_{(j-1):(p-2)}^T \Big)^T = \boldsymbol{\theta}_j\tau_{0j}^2 = \boldsymbol{\phi}_{0j}$. We thus have that $||\boldsymbol{\theta}_j||_0 = ||\boldsymbol{d}_{0j}|| + 1 \leqslant \widetilde{s} + 1$.

To prove the second part of the lemma, note that $\mathrm{E}(\widetilde{\boldsymbol{x}}_{-1}\widetilde{\boldsymbol{x}}_{-1}^T) = \mathrm{E}[\mathrm{Cov}(\boldsymbol{x}_{-1}|\boldsymbol{x}^T\boldsymbol{\beta}_0)]$. Assumption (A2) indicates that $\sup_{||\boldsymbol{v}||_2=1} \boldsymbol{v}^T\mathrm{E}(\widetilde{\boldsymbol{x}}_{-1}\widetilde{\boldsymbol{x}}_{-1}^T)\boldsymbol{v} = \xi_1$. We can derive that

$$\sup_{||\boldsymbol{v}||_2=1} \boldsymbol{v}^T\boldsymbol{\Omega}\boldsymbol{v} = \sup_{||\boldsymbol{v}||_2=1} \mathrm{E}\big\{[G^{(1)}(\boldsymbol{x}^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)(\widetilde{\boldsymbol{x}}_{-1}^T\boldsymbol{v})]^2\big\} \leqslant b^2 \sup_{||\boldsymbol{v}||_2=1} \boldsymbol{v}^T\mathrm{E}(\widetilde{\boldsymbol{x}}_{-1}\widetilde{\boldsymbol{x}}_{-1}^T)\boldsymbol{v} = b^2\xi_1.$$

Recall that $\tau_{0j}^2 = \Omega_{(j-1),(j-1)} - (\boldsymbol{\Omega}_{-(j-1),(j-1)})^T(\boldsymbol{\Omega}_{-(j-1),-(j-1)})^{-1}\boldsymbol{\Omega}_{-(j-1),(j-1)}$. Since $\boldsymbol{\Omega}$ is positive definite, we thus have that $(\boldsymbol{\Omega}_{-(j-1),-(j-1)})^{-1}$ is positive definite, and $\tau_{0j}^2 \leqslant \Omega_{(j-1),(j-1)} = \boldsymbol{e}_{j-1}^T\boldsymbol{\Omega}\boldsymbol{e}_{j-1} \leqslant b^2\xi_1$, uniformly in $j = 2, \cdots, p$.

Since $\boldsymbol{\Theta} = \boldsymbol{\Omega}^{-1}$, we know that $\boldsymbol{\theta}_j^T\boldsymbol{\Omega}\boldsymbol{\theta}_j = \Theta_{(j-1),(j-1)}$. We observe that $\tau_{0j}^{-2} = \Theta_{(j-1),(j-1)} \leqslant ||\boldsymbol{\theta}_j||_2$. Note that $\tau_{0j}^{-2} = \Omega_{(j-1),(j-1)}^{-1} \leqslant \xi_2^{-1}$ by Assumption (A2)-(a). We observe that $||\boldsymbol{\theta}_j||_2 \geqslant \boldsymbol{\theta}_j^T\boldsymbol{\Omega}\boldsymbol{\theta}_j = \tau_{0j}^{-4}(\boldsymbol{\phi}_{0j}^T\boldsymbol{\Omega}\boldsymbol{\phi}_{0j}) \geqslant \xi_2\tau_{0j}^{-4}||\boldsymbol{\phi}_{0j}||_2^2 = \xi_2||\boldsymbol{\theta}_j||_2^2$, where the second inequality applies Assumption (A2)-(a). It implies that $\tau_{0j}^{-2} \leqslant ||\boldsymbol{\theta}_j||_2 \leqslant \xi_2^{-1}$ uniformly in j, which completes the proof of the lemma. $\square$

Proof of Lemma A14. It suffices to show that

$$\max_{2\leqslant j,k\leqslant p}\left|\frac{1}{n}\sum_{i=1}^{n}[\widetilde{Y}_i-\widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2\widehat{\boldsymbol{\theta}}_j^T\widehat{\boldsymbol{x}}_{i,-1}\widehat{\boldsymbol{x}}_{i,-1}^T\widehat{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_j^T\boldsymbol{\Lambda}\boldsymbol{\theta}_k\right|=o_p(1).$$

Rewrite that

$$\begin{aligned}
&\max_{2\leqslant j,k\leqslant p}\left|\frac{1}{n}\sum_{i=1}^{n}[\widetilde{Y}_i-\widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2\widehat{\boldsymbol{\theta}}_j^T\widehat{\boldsymbol{x}}_{i,-1}\widehat{\boldsymbol{x}}_{i,-1}^T\widehat{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_j^T\boldsymbol{\Lambda}\boldsymbol{\theta}_k\right|\\
\leqslant&\max_{2\leqslant j,k\leqslant p}\left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{\epsilon}_i^2[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2\widehat{\boldsymbol{\theta}}_j^T(\widehat{\boldsymbol{x}}_{i,-1}\widehat{\boldsymbol{x}}_{i,-1}^T-\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T)\widehat{\boldsymbol{\theta}}_k\right|\\
&+\max_{2\leqslant j,k\leqslant p}\left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{\epsilon}_i^2\{[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2-[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]^2\}\widehat{\boldsymbol{\theta}}_j^T\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T\widehat{\boldsymbol{\theta}}_k\right|\\
&+\max_{2\leqslant j,k\leqslant p}\left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{\epsilon}_i^2[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]^2\widehat{\boldsymbol{\theta}}_j^T\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T\widehat{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_j^T\boldsymbol{\Lambda}\boldsymbol{\theta}_k\right|\\
&+\max_{2\leqslant j,k\leqslant p}\left|\frac{1}{n}\sum_{i=1}^{n}[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)-\widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2\widehat{\boldsymbol{\theta}}_j^T\widehat{\boldsymbol{x}}_{i,-1}\widehat{\boldsymbol{x}}_{i,-1}^T\widehat{\boldsymbol{\theta}}_k\right|\\
&+\max_{2\leqslant j,k\leqslant p}\left|\frac{2}{n}\sum_{i=1}^{n}\widetilde{\epsilon}_i[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)-\widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})][\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2\widehat{\boldsymbol{\theta}}_j^T\widehat{\boldsymbol{x}}_{i,-1}\widehat{\boldsymbol{x}}_{i,-1}^T\widehat{\boldsymbol{\theta}}_k\right|\\
\triangleq&\sum_{l=1}^{5}\max_{2\leqslant j,k\leqslant p}|J_{njkl}|,
\end{aligned}$$

where J_{njkl}'s are defined in the context. Note that

$$\begin{aligned}
\max_{2\leqslant j,k\leqslant p}|J_{njk1}|\leqslant&\max_{2\leqslant j,k\leqslant p}\left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{\epsilon}_i^2[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2\boldsymbol{\theta}_j^T(\widehat{\boldsymbol{x}}_{i,-1}\widehat{\boldsymbol{x}}_{i,-1}^T-\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T)\boldsymbol{\theta}_k\right|\\
&+\max_{2\leqslant j,k\leqslant p}\left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{\epsilon}_i^2[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2(\widehat{\boldsymbol{\theta}}_j-\boldsymbol{\theta}_j)^T(\widehat{\boldsymbol{x}}_{i,-1}\widehat{\boldsymbol{x}}_{i,-1}^T-\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T)\boldsymbol{\theta}_k\right|\\
&+\max_{2\leqslant j,k\leqslant p}\left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{\epsilon}_i^2[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2\boldsymbol{\theta}_j^T(\widehat{\boldsymbol{x}}_{i,-1}\widehat{\boldsymbol{x}}_{i,-1}^T-\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T)(\widehat{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k)\right|\\
&+\max_{2\leqslant j,k\leqslant p}\left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{\epsilon}_i^2[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2(\widehat{\boldsymbol{\theta}}_j-\boldsymbol{\theta}_j)^T(\widehat{\boldsymbol{x}}_{i,-1}\widehat{\boldsymbol{x}}_{i,-1}^T-\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T)(\widehat{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k)\right|\\
\triangleq&\max_{2\leqslant j,k\leqslant p}|J_{njk11}|+\max_{2\leqslant j,k\leqslant p}|J_{njk12}|+\max_{2\leqslant j,k\leqslant p}|J_{njk13}|+\max_{2\leqslant j,k\leqslant p}|J_{njk14}|.
\end{aligned}$$

Consider the event $\mathcal{F}_n = \{\max_{1\leqslant i\leqslant n} |\widetilde{\epsilon}_i| \leqslant (\sigma_\epsilon + M)\sqrt{\log(p \vee n)}\}$, which holds with probability at least $1 - \exp[-c\log(p \vee n)]$ by the sub-Gaussian property for $\widetilde{\epsilon}$, , for some constants $c > 0$, and all n sufficiently large. Then by the proofs of Lemma B10, we can derive that

$$\max_{2\leqslant j,k\leqslant p} |J_{njk11}| \leqslant O_p(\log(p \vee n)) * O_p(\sqrt{s}h^2) = O_p(\sqrt{s}h^2 \log(p \vee n)) = o_p(nh^7) = o_p(1),$$

with probability at least $1 - \exp(-c_1 \log p)$, for some constant $c_1 > 0$, and all n sufficiently large. Similarly, we have that $\max_{2\leqslant j,k\leqslant p} |J_{njk12}| \leqslant O_p(\sqrt{s}h^2 \log(p \vee n)) * \max_{2\leqslant j\leqslant p} ||\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j||_1 = o_p(nh^7) * O_p(\widetilde{s}\eta) = o_p(1)$, and $\max_{2\leqslant j,k\leqslant p} |J_{njk13}| = o_p(1)$, with the same probability bound. Finally, we can derive that $\max_{2\leqslant j,k\leqslant p} |J_{njk14}| \leqslant O_p(\sqrt{s}h^2 \log(p\vee n)) * \max_{2\leqslant j\leqslant p} ||\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j||_1^2 = o_p(nh^7) * O_p(\widetilde{s}^2\eta^2) = o_p(1)$.

To bound $\max_{2\leqslant j,k\leqslant p} |J_{njk2}|$, Lemma A11 and Lemma B3 together imply that

$$\begin{aligned}
&\max_{2\leqslant j,k\leqslant p} |J_{njk2}| \\
\leqslant &\max_{1\leqslant i\leqslant n} \left|[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2 - [G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]^2\right| * \left[\frac{1}{n}\sum_{i=1}^n \widetilde{\epsilon}_i^6\right]^{1/3} * \max_{2\leqslant j,k\leqslant p} \left[\frac{1}{n}\sum_{i=1}^n |\widehat{\boldsymbol{\theta}}_j^T \widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T\widehat{\boldsymbol{\theta}}_k|^{3/2}\right]^{2/3} \\
\leqslant &O_p(h) * \left[\frac{1}{n}\sum_{i=1}^n \widetilde{\epsilon}_i^6\right]^{1/3} * \max_{2\leqslant j\leqslant p} \left[\frac{1}{n}\sum_{i=1}^n |\widetilde{\boldsymbol{x}}_{i,-1}^T\widehat{\boldsymbol{\theta}}_j|^3\right]^{2/3} \\
\leqslant &O_p(h) * O_p(1) = O_p(h) = o_p(1),
\end{aligned}$$

with probability at least $1 - \exp[-c_1\log(p \wedge n)]$, for some positive constant c_1, and all n sufficiently large. We can also derive that

$$\begin{aligned}
\max_{2\leqslant j,k\leqslant p} |J_{njk4}| &\leqslant \max_{1\leqslant i\leqslant n} [\widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})] - G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]^2 * \max_{1\leqslant i\leqslant n} [\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2 * \max_{2\leqslant j\leqslant n} \frac{1}{n}\sum_{i=1}^n (\widehat{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j)^2 \\
&\leqslant O_p(s^2h^4\log(p \vee n)) * O_p(1) * O_p(1) = O_p(s^2h^4\log(p\vee n)) = o_p(1),
\end{aligned}$$

with probability at least $1 - \exp(-c_1 \log p)$for some positive constant c_1, and all n sufficiently large, by Lemma B11, Lemma A5 and Lemma A11. Conditional on the event $\mathcal{F}_n$, we can see $\max_{2\leqslant j,k\leqslant p}|J_{njk5}| \leqslant O_p(sh^2\log(p\vee n)) = o_p(1)$. To bound $\max_{2\leqslant j,k\leqslant p}|J_{njk4}|$, let $\widehat{\boldsymbol{\Lambda}} = \frac{1}{n}\sum_{i=1}^n \widetilde{\epsilon}_i^2[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]^2\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T$, then we can rewrite it as

$$\begin{aligned}\max_{2\leqslant j,k\leqslant p}|J_{njk4}| &\leqslant \max_{2\leqslant j,k\leqslant p}\left|\boldsymbol{\theta}_j^T\{\widehat{\boldsymbol{\Lambda}}-\boldsymbol{\Lambda}\}\boldsymbol{\theta}_k\right| + \max_{2\leqslant j,k\leqslant p}\left|(\widehat{\boldsymbol{\theta}}_j-\boldsymbol{\theta}_j)^T\widehat{\boldsymbol{\Lambda}}\boldsymbol{\theta}_k\right| + \max_{2\leqslant j,k\leqslant p}\left|\widehat{\boldsymbol{\theta}}_j^T\widehat{\boldsymbol{\Lambda}}(\widehat{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k)\right| \\ &= \max_{2\leqslant j,k\leqslant p}|J_{njk41}| + \max_{2\leqslant j,k\leqslant p}|J_{njk42}| + \max_{2\leqslant j,k\leqslant p}|J_{njk43}|,\end{aligned}$$

where the definition of J_{njk4l} is clear. Given $(\epsilon_i, \boldsymbol{x}_i)$, for any $\boldsymbol{\theta}_j$, $\widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j$ is sub-Gaussian with variance proxy at most $b^2[\epsilon_i+g(\boldsymbol{x}_i)]^2(\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j)^2$. Hence, we can conclude that $\frac{1}{n}\sum_{i=1}^n \widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j$ is sub-Gaussian with variance proxy at most $\frac{b^2}{n}\sum_{i=1}^n[\epsilon_i + g(\boldsymbol{x}_i)]^2(\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\theta}_j)^2 \xrightarrow{p} c$ for constant $c>0$. Lemma B1 implies that

$$P\left(\max_{2\leqslant j,k\leqslant p}|J_{njk41}| \geqslant c\sqrt{\frac{\log p}{n}}\right) \leqslant \sum_{j,k} P\left(|J_{njk41}| \geqslant c\sqrt{\frac{\log p}{n}}\right) \leqslant p^2\exp(-c_1\log p),$$

for some positive constant c_1, and all n sufficiently large, where $\sqrt{\frac{\log p}{n}} = o(h^{5/2})$. For J_{njk42}, we can conclude that

$$\begin{aligned}\max_{2\leqslant j,k\leqslant p}|J_{njk42}| &\leqslant b^2\left[\frac{1}{n}\sum_{i=1}^n \widetilde{\epsilon}_i^6\right]^{1/3} * \max_{1\leqslant k\leqslant p}\left[\frac{1}{n}\sum_{i=1}^n|\widetilde{\boldsymbol{x}}_{i,-1}^T\widehat{\boldsymbol{\theta}}_k|^3\right]^{1/3} * \max_{2\leqslant j\leqslant p}\left[\frac{1}{n}\sum_{i=1}^n|\widetilde{\boldsymbol{x}}_{i,-1}^T(\widehat{\boldsymbol{\theta}}_j-\boldsymbol{\theta}_j)|^3\right]^{1/3} \\ &\leqslant O_p(\widetilde{s}^{1/2}h) = o_p(1).\end{aligned}$$

Similar proof works to bound $\max_{2\leqslant j,k\leqslant p}|J_{njk43}|$. The lemma is proved. □

S7 Auxiliary results

Lemma B1 (Lemma 14 in Loh and Wainwright [2012])**.** *Let p_1, p_2 be two arbitrary positive integers. If $\{\boldsymbol{x}_i \in \mathbb{R}^{p_1} : i = 1,\cdots,n\}$ are independent zero-mean sub-Gaussian random* *vectors with variance proxy σ_x^2, then for any fixed unit vector $\boldsymbol{v} \in \mathbb{R}^{p_1}$, $\forall\ t > 0$,*

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left[(\boldsymbol{x}_i^T\boldsymbol{v})^2 - \mathrm{E}(\boldsymbol{x}_i^T\boldsymbol{v})^2\right]\right| \geqslant t\right) \leqslant 2\exp\left[-cn\min\left(\frac{t^2}{\sigma_x^4}, \frac{t}{\sigma_x^2}\right)\right], \tag{S45}$$

for some universal constant $c > 0$. Moreover, if $\{\boldsymbol{y}_i \in \mathbb{R}^{p_2} : i = 1, \cdots, n\}$ are independent zero-mean sub-Gaussian random vectors with variance proxy σ_y^2, then $\forall\ t > 0$,

$$P\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\left[\boldsymbol{x}_i\boldsymbol{y}_i^T - \mathrm{E}(\boldsymbol{x}_i\boldsymbol{y}_i^T)\right]\right\|_\infty \geqslant t\right) \leqslant 6p_1p_2\exp\left[-cn\min\left(\frac{t^2}{\sigma_x^2\sigma_y^2}, \frac{t}{\sigma_x\sigma_y}\right)\right]. \tag{S46}$$

Let if $p = p_1 \vee p_2$. If $\log p = O(n)$, then there exist universal positive constants c_0, c_1 and c_2 such that

$$P\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\left[\boldsymbol{x}_i\boldsymbol{y}_i^T - \mathrm{E}(\boldsymbol{x}_i\boldsymbol{y}_i^T)\right]\right\|_\infty \geqslant c_0\sigma_x\sigma_y\sqrt{\frac{\log p}{n}}\right) \leqslant c_1\exp(-c_2\log p). \tag{S47}$$

Lemma B2 (Lemma 15 in Loh and Wainwright [2012])**.** *Let $\mathbb{K}(s_0) = \{\boldsymbol{v} \in \mathbb{R}^p : ||\boldsymbol{v}||_2 \leqslant 1, ||\boldsymbol{v}||_0 \leqslant s_0\}$. If $\{\boldsymbol{x}_i \in \mathbb{R}^p : i = 1, \cdots, n\}$ are independent zero-mean sub-Gaussian random vectors with variance proxy σ_x^2, then there is a universal constant $c > 0$ such that for any $s_0 \geqslant 1$,*

$$P\Big(\sup_{\boldsymbol{v}\in\mathbb{K}(2s_0)}\left|\frac{1}{n}\sum_{i=1}^{n}\left[(\boldsymbol{x}_i^T\boldsymbol{v})^2 - \mathrm{E}(\boldsymbol{x}_i^T\boldsymbol{v})^2\right]\right| \geqslant t\Big) \leqslant 2\exp\left[-cn\min\left(\frac{t^2}{\sigma_x^4}, \frac{t}{\sigma_x^2}\right) + 2s_0\log p\right].$$

Lemma B3. *Let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ be independent sub-Gaussian random vectors with variance proxy σ_x^2. For any $s_0 \geqslant 1$, there exists a universal constant $c > 0$ such that for all n sufficiently large.*

$$P\Big(\sup_{\boldsymbol{\gamma}\in\mathbb{K}(p,2s_0)}\left|\frac{1}{n}\sum_{i=1}^{n}\left[|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^3 - \mathrm{E}|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^3\right]\right| \geqslant t\Big) \leqslant \exp\left\{-c\min\left[\frac{nt^2}{\sigma_x^6}, \frac{(nt)^{2/3}}{\sigma_x^2}\right] + 2s_0\log p\right\};$$
$$P\Big(\sup_{\boldsymbol{\gamma}\in\mathbb{K}(p,2s_0)}\left|\frac{1}{n}\sum_{i=1}^{n}\left[|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^4 - \mathrm{E}|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^4\right]\right| \geqslant t\Big) \leqslant \exp\left\{-c\min\left[\frac{nt^2}{\sigma_x^8}, \frac{(nt)^{1/2}}{\sigma_x^2}\right] + 2s_0\log p\right\}.$$

Proof. For any fixed $\boldsymbol{\gamma} \in \mathbb{R}^p$ such that $||\boldsymbol{\gamma}||_2 \leqslant 1$, $\boldsymbol{x}_i^T\boldsymbol{\gamma}$ is also sub-Gaussian with variance proxy bounded by σ^2. Applying the result on concentration inequality for the polynomial functions of independent sub-Gaussian random variables, Theorem 1.4 of Adamczak and Wolff [2015] and the example in their section 3.1.2, we have $\forall\ t > 0$, there exist universal positive constants c_1 and c_2 such that

$$P\Big(\frac{1}{n}\sum_{i=1}^{n}\big(|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^3 - \mathrm{E}|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^3\big) \geqslant t\Big) \leqslant c_1 \exp\left\{-c_2 \min\left[\frac{nt^2}{\sigma_x^6}, \frac{(nt)^{2/3}}{\sigma_x^2}\right]\right\}, \tag{S48}$$

$$P\Big(\frac{1}{n}\sum_{i=1}^{n}\big(|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^4 - \mathrm{E}|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^4\big) \geqslant t\Big) \leqslant c_1 \exp\left\{-c_2 \min\left[\frac{nt^2}{\sigma_x^8}, \frac{(nt)^{1/2}}{\sigma_x^2}\right]\right\}. \tag{S49}$$

Next, we apply the covering technique of Lemma B2 to extend the above probability bound to uniformly on $\mathbb{K}(p, 2s_0) = \{\boldsymbol{v} \in \mathbb{R}^p : ||\boldsymbol{v}||_2 \leqslant 1, ||\boldsymbol{v}||_0 \leqslant 2s_0\}$.

For any subset $\mathcal{U} \subseteq \{1, \cdots, p\}$, define $\mathcal{S}_{\mathcal{U}} = \{\boldsymbol{\gamma} \in \mathbb{R}^p : ||\boldsymbol{\gamma}||_2 \leqslant 1, \mathrm{supp}(\boldsymbol{\gamma}) \subseteq \mathcal{U}\}$. Then $\mathbb{K}(p, 2s_0) = \bigcup_{|\mathcal{U}| \leqslant 2s_0} \mathcal{S}_{\mathcal{U}}$. Let $\mathcal{A} = \{u_1, \cdots, u_m\}$ be a $\frac{1}{4}-$cover of $\mathcal{S}_{\mathcal{U}}$, that is $\forall\ \boldsymbol{\gamma} \in \mathcal{S}_{\mathcal{U}}$, there exists some $\boldsymbol{\xi} \in \mathcal{A}$ such that $||\boldsymbol{\gamma} - \boldsymbol{\xi}||_2 \leqslant 1/4$. We can construct $\mathcal{A}$ such that $|\mathcal{A}| \leqslant 16^{2s_0}$. We observe that

$$\begin{aligned}\Big|\frac{1}{n}\sum_{i=1}^{n}\big(|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^3 - |\boldsymbol{x}_i^T\boldsymbol{\xi}|^3\big)\Big| \leqslant & \frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^2|\boldsymbol{x}_i^T(\boldsymbol{\gamma}-\boldsymbol{\xi})| + \frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{x}_i^T\boldsymbol{\gamma}| * |\boldsymbol{x}_i^T\boldsymbol{\xi}| * |\boldsymbol{x}_i^T(\boldsymbol{\gamma}-\boldsymbol{\xi})| \\ & + \frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{x}_i^T\boldsymbol{\xi}|^2|\boldsymbol{x}_i^T(\boldsymbol{\gamma}-\boldsymbol{\xi})|.\end{aligned}$$

By Hölder inequality,

$$\frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^2|\boldsymbol{x}_i^T(\boldsymbol{\gamma}-\boldsymbol{\xi})| \leqslant \Big(\frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^3\Big)^{2/3} * \Big(\frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{x}_i^T(\boldsymbol{\gamma}-\boldsymbol{\xi})|^3\Big)^{1/3}.$$

Since $4(\boldsymbol{\gamma} - \boldsymbol{\xi}) \in \mathcal{S}_{\mathcal{U}}$, we have

$$\sup_{\boldsymbol{\gamma}\in\mathcal{S}_{\mathcal{U}}}\sup_{\boldsymbol{\xi}\in\mathcal{A}}\frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^2|\boldsymbol{x}_i^T(\boldsymbol{\gamma}-\boldsymbol{\xi})| \leqslant \frac{1}{4n}\sup_{\boldsymbol{\gamma}\in\mathcal{S}_{\mathcal{U}}}\sum_{i=1}^{n}|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^3.$$

Similarly analysis applies to the other two terms. Note that $\max_{\boldsymbol{\xi}\in\mathcal{A}} \frac{1}{n}\sum_{i=1}^n |\boldsymbol{x}_i^T\boldsymbol{\xi}|^3 \leqslant \sup_{\boldsymbol{\gamma}\in\mathcal{S}_\mathcal{U}}$ $\frac{1}{n}\sum_{i=1}^n |\boldsymbol{x}_i^T\boldsymbol{\gamma}|^3$, then we have

$$\sup_{\boldsymbol{\gamma}\in\mathcal{S}_\mathcal{U}} \frac{1}{n}\sum_{i=1}^n |\boldsymbol{x}_i^T\boldsymbol{\gamma}|^3 \leqslant \max_{\boldsymbol{\xi}\in\mathcal{A}} \frac{1}{n}\sum_{i=1}^n |\boldsymbol{x}_i^T\boldsymbol{\xi}|^3 + \frac{3}{4n}\sup_{\boldsymbol{\gamma}\in\mathcal{S}_\mathcal{U}} \sum_{i=1}^n |\boldsymbol{x}_i^T\boldsymbol{\gamma}|^3,$$

which implies that $\sup_{\boldsymbol{\gamma}\in\mathcal{S}_\mathcal{U}} \frac{1}{n}\sum_{i=1}^n |\boldsymbol{x}_i^T\boldsymbol{\gamma}|^3 \leqslant 4\max_{\boldsymbol{\xi}\in\mathcal{A}} \frac{1}{n}\sum_{i=1}^n |\boldsymbol{x}_i^T\boldsymbol{\xi}|^3$. Combining this with (S48) and applying the union bound, there exists a universal constant $c > 0$ such that

$$\begin{aligned} P\Big(\sup_{\boldsymbol{\gamma}\in\mathcal{S}_\mathcal{U}} \frac{1}{n}\sum_{i=1}^n \big(|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^3 - \mathrm{E}|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^3\big) \geqslant 4t\Big) &\leqslant P\Big(\max_{\boldsymbol{\xi}\in\mathcal{A}} \frac{1}{n}\sum_{i=1}^n \big(|\boldsymbol{x}_i^T\boldsymbol{\xi}|^3 - \mathrm{E}|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^3\big) \geqslant t\Big) \\ &\leqslant 16^{2s_0}\exp\left\{-c\min\left[\frac{nt^2}{\sigma_x^6}, \frac{(nt)^{2/3}}{\sigma_x^2}\right]\right\}. \end{aligned}$$

Taking a union bound over the $\binom{p}{2s_0} \leqslant p^{2s_0}$ choices of $\mathcal{U}$ for $\mathbb{K}(p, 2s_0)$ yields that for all n sufficiently large,

$$P\Big(\sup_{\boldsymbol{\gamma}\in\mathbb{K}(p,2s_0)} \Big|\frac{1}{n}\sum_{i=1}^n \big(|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^3 - \mathrm{E}|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^3\big)\Big| \geqslant t\Big) \leqslant \exp\left\{-c\min\left[\frac{nt^2}{\sigma_x^6}, \frac{(nt)^{2/3}}{\sigma_x^2}\right] + 2s_0\log p\right\}.$$

Hence, the first claim of Lemma B3 is proved. The second claim can be probed similarly. □

Lemma B4. *Under the assumptions of Theorem 1, there exist some positive constants c_0, c_1 such that for all n sufficiently large,*

$$P\left(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}} |A_{n1}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})| \geqslant c_0 h^2\right) \leqslant \exp(-c_1 n h^5).$$

Proof. Write $a_{nj}(t|\boldsymbol{\beta}) = K\Big(\frac{t-\boldsymbol{x}_j^T\boldsymbol{\beta}}{h}\Big)\widetilde{\epsilon}_j$. Then $A_{n1}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) = [(n-1)h]^{-1}\sum_{j=1,j\neq i}^n a_{nj}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})$. Let $f_{\boldsymbol{\beta}}(\cdot)$ denote the p.d.f of $\boldsymbol{x}^T\boldsymbol{\beta}$. Assumption (A1), (A3) and (A4) together imply that

$$\mathrm{E}[a_{nj}^2(t|\boldsymbol{\beta})] \leqslant (\sigma_\epsilon^2 + M^2)\mathrm{E}\Big\{K^2\Big(\frac{t-\boldsymbol{x}_j^T\boldsymbol{\beta}}{h}\Big)\Big\} = (\sigma_\epsilon^2 + M^2)h\int K^2(z)f_{\boldsymbol{\beta}}(t-hz)dz \leqslant ch,$$

for some positive constant c. Note that $\mathrm{E}[a_{nj}(t|\boldsymbol{\beta})] = 0$. Note that ϵ_i is sub-Gaussian, $K(\cdot)$ and $g(\cdot)$ are bounded almost everywhere. It is easy to see that $\mathrm{E}\big[|a_{nj}(t|\boldsymbol{\beta})|^k\big] \leqslant \frac{1}{2}\mathrm{E}[a_{nj}^2(t|\boldsymbol{\beta})]L^{k-2}k!$, for some positive real L and every integer $k \geqslant 2$. For any fixed $\boldsymbol{\beta}$ and $0 \leqslant v \leqslant \frac{1}{2L}\sqrt{(n-1)\mathrm{E}[a_{nj}^2(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})]}$, by Bernstein's inequality,

$$P\left(\Big|\sum_{j=1,j\neq i}^{n} a_{nj}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\Big| \geqslant 2v\sqrt{c(n-1)h}\right) \leqslant 2\exp(-v^2).$$

Taking $v = \sqrt{c(n-1)h^5}$, we have

$$P\left(\big|A_{n1}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big| \geqslant ch^2\right) \leqslant 2\exp(-cnh^5).$$

To obtain the uniform bound, we cover $\mathbb{B}$ with L_2-balls with radius δ. Denote the unit Euclidean sphere as $\mathcal{S}^{ks} = \{\boldsymbol{v} \in \mathbb{R}^{ks} : ||\boldsymbol{v}||_2 = 1\}$. Let the covering number $N(\delta, \mathcal{S}^{ks}, \rho)$ be the minimum n such that there exists an $\delta-$cover of $\mathcal{S}^{ks}$ of size n, with respect to the L_2 distance ρ. It is well known that $N(\delta, \mathcal{S}^{ks}, \rho) \leqslant (1 + \frac{2}{\delta})^{ks}$. Consider the decomposition

$$\big\{\boldsymbol{\beta} \in \mathcal{S}^p : ||\boldsymbol{\beta}||_0 = ks\big\} = \bigcup_{\mathcal{S}\subseteq[p]:|\mathcal{S}|=ks} \big\{\boldsymbol{\beta} \in \mathcal{S}^p : \mathrm{supp}(\boldsymbol{\beta}) = \mathcal{S}\big\},$$

where $|\mathcal{S}|$ is the cardinal number of $\mathcal{S}$. Let $\mathcal{N}_\delta$ be an $\delta-$cover of $\mathbb{B} = \{\boldsymbol{\beta} \in \mathbb{B}_0 : ||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_2 \leqslant r, ||\boldsymbol{\beta}||_0 \leqslant ks\}$. it is easy to show the covering number $N = |\mathcal{N}_\delta|$ satisfies

$$N \leqslant \left\{\begin{pmatrix} p \\ ks \end{pmatrix}\Big(1 + \frac{2r}{\delta}\Big)^{ks}\right\}^2 \leqslant \left\{\Big(1 + \frac{2r}{\delta}\Big)\frac{ep}{ks}\right\}^{2ks} \leqslant c_2\Big(\frac{p}{\delta}\Big)^{2ks},$$

for sufficiently large c_2. For any $\boldsymbol{\beta}$ in such a ball with center $\boldsymbol{\beta}^*$, let us take $\delta = \frac{h^4}{2\sqrt{n}}$, then the Lipschitz condition for $K(\cdot)$ implies that

$$
\begin{aligned}
&\left|(n-1)^{-1}\sum_{j=1,j\neq i}^{n}[K_h(\boldsymbol{x}_i^T\boldsymbol{\beta}-\boldsymbol{x}_j^T\boldsymbol{\beta})-K_h(\boldsymbol{x}_i^T\boldsymbol{\beta}^*-\boldsymbol{x}_j^T\boldsymbol{\beta}^*)]\widetilde{\epsilon}_j\right| \\
\leqslant& c_0[(n-1)h^2]^{-1}\sum_{j=1,j\neq i}^{n}|(\boldsymbol{x}_i-\boldsymbol{x}_j)^T(\boldsymbol{\beta}-\boldsymbol{\beta}^*)|*|\widetilde{\epsilon}_j| \\
\leqslant& 2c_0h^{-2}\sqrt{|\boldsymbol{x}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}^*)|^2+(n-1)^{-1}\sum_{j=1,j\neq i}^{n}|\boldsymbol{x}_j^T(\boldsymbol{\beta}-\boldsymbol{\beta}^*)|^2}*\sqrt{(n-1)^{-1}\sum_{j=1,j\neq i}^{n}\widetilde{\epsilon}_j^2},
\end{aligned}
$$

for some positive constant c_0. Lemma A2 and Lemma B1 imply that $P\Big(\Big|(n-1)^{-1}\sum_{j=1,j\neq i}^{n}\widetilde{\epsilon}_j^2-4(\sigma_\epsilon^2+M^2)\Big|\geqslant\sigma_\epsilon^2+M^2\Big)\leqslant\exp(-c_1n)$ for some constant $c_1>0$. Lemma B2 suggests

$$
P\left(\left|(n-1)^{-1}\sum_{j=1,j\neq i}^{n}|\boldsymbol{x}_j^T(\boldsymbol{\beta}-\boldsymbol{\beta}^*)|^2-(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^T\mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T)(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\right|\geqslant\sigma_x^2||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2^2,\ \forall\ \boldsymbol{\beta},\boldsymbol{\beta}^*\in\mathbb{B}\right)\leqslant\exp(-c_1n),
$$

for some constant $c_1>0$ and all n sufficiently large. Taking $t=(n-1)\sigma_x^2$ and $s_0=ks$, Lemma B2 suggests

$$
P\left(|\boldsymbol{x}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}^*)|\geqslant\sqrt{n}\sigma_x||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2,\ \forall\ \boldsymbol{\beta},\boldsymbol{\beta}^*\in\mathbb{B}\right)\leqslant\exp(-c_1n),
$$

for some positive constant c_1 and all n sufficiently large. Define the event

$$
\begin{aligned}
\mathcal{E}_1=&\left\{(n-1)^{-1}\sum_{j=1,j\neq i}^{n}\widetilde{\epsilon}_j^2\leqslant5(\sigma_\epsilon^2+M^2),\ |\boldsymbol{x}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}^*)|\geqslant\sqrt{n}\sigma_x^2||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2,\ \forall\ \boldsymbol{\beta},\boldsymbol{\beta}^*\in\mathbb{B}\right\} \\
&\bigcap\left\{(n-1)^{-1}\sum_{j=1,j\neq i}^{n}|\boldsymbol{x}_j^T(\boldsymbol{\beta}-\boldsymbol{\beta}^*)|^2\leqslant(\xi_3+\sigma_x^2)||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2^2,\ \forall\ \boldsymbol{\beta},\boldsymbol{\beta}^*\in\mathbb{B}\right\}.
\end{aligned}
$$

We have that $P(\mathcal{E}_1)\geqslant1-3\exp(-c_1n)$ for all n sufficiently large, according to the above discussions. Hence on the $\mathcal{E}_1$, we have

$$
\begin{aligned}
&\left|(n-1)^{-1}\sum_{j=1,j\neq i}^{n}[K_h(\boldsymbol{x}_i^T\boldsymbol{\beta}-\boldsymbol{x}_j^T\boldsymbol{\beta})-K_h(\boldsymbol{x}_i^T\boldsymbol{\beta}^*-\boldsymbol{x}_j^T\boldsymbol{\beta}^*)]\widetilde{\epsilon}_j\right| \\
\leqslant & 2c_0h^{-2}\sqrt{5(\sigma_\epsilon^2+M^2)}*\sqrt{n\sigma_x^2+\sigma_x^2+\xi_3}*||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2 \\
\leqslant & c_2h^{-2}\sqrt{n}||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2\leqslant c_2h^{-2}\sqrt{n}\delta=c_2h^2/2,
\end{aligned}
$$

for some positive constant c_2 and all n sufficiently large. We thus have

$$
\begin{aligned}
P\left(\sup_{\boldsymbol{\beta}\in\mathbb{B}}|A_{n1}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})|\geqslant c_2h^2\right)\leqslant & P\left(\bigcup_{\boldsymbol{\beta}^*\in\mathcal{N}_\delta}|A_{n1}(\boldsymbol{x}_i^T\boldsymbol{\beta}^*|\boldsymbol{\beta}^*)|\geqslant c_2h^2/2\right) \\
& +P\left(\sup_{\boldsymbol{\beta}^*\in\mathcal{N}_\delta}\sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2\leqslant\delta}|A_{n1}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-A_{n1}(\boldsymbol{x}_i^T\boldsymbol{\beta}^*|\boldsymbol{\beta}^*)|\leqslant c_2h^2/2\right) \\
\leqslant & \sum_{\boldsymbol{\beta}^*\in\mathcal{N}_\delta}P\left(|A_{n1}(\boldsymbol{x}_i^T\boldsymbol{\beta}^*|\boldsymbol{\beta}^*)|\geqslant c_2h^2/2\right)+1-P(\mathcal{E}_1) \\
\leqslant & cp^{2ks}\delta^{-2ks}\exp(-c_2nh^5)+3\exp(-c_1n) \\
= & c\exp\left(2ks\log p-2ks\log\delta-c_2nh^5\right)+3\exp(-c_1n).
\end{aligned}
$$

By the assumptions of Theorem 1, $d_0s\log(p\vee n)\leqslant nh^5$ for some constant d_0. Thus $-s\log\delta=c*s\log(h^{-1})+c*s\log n\leqslant s\log(p\vee n)$. It is followed that

$$
P\left(\sup_{\boldsymbol{\beta}\in\mathbb{B}}|A_{n1}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})|\geqslant c_0h^2\right)\leqslant\exp(-cnh^5),
$$

for some positive constants c_0 and c. We therefore have

$$
\begin{aligned}
P\left(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}}|A_{n1}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})|\geqslant c_0h^2\right)\leqslant & \sum_{i=1}^{n}P\left(\sup_{\boldsymbol{\beta}\in\mathbb{B}}|A_{n1}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})|\geqslant c_0h^2\right) \\
\leqslant & n\exp(-cnh^5)=\exp[-(cnh^5-\log n)] \\
\leqslant & \exp(-c_1nh^5),
\end{aligned}
$$

for some positive constants c_0 c_1, and all n sufficiently large. $\square$

Lemma B5. *Under the assumptions of Theorem 1, there exist some positive constants c_0, c_1 such that for all n sufficiently large,*

$$P\left(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}}|A_{n2}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})|\geqslant c_0h^2\right)\leqslant\exp[-c_1\log(p\vee n)].$$

Proof. Let $A_{n2}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})=[(n-1)h]^{-1}\sum_{j=1,j\neq i}^n\gamma_i(z_j)$, where $\gamma_i(z_j)=K\Big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}-\boldsymbol{x}_j^T\boldsymbol{\beta}}{h}\Big)\big[f_0(\boldsymbol{x}_j^T\boldsymbol{\beta}_0)-G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]$. Recall that $f_{\boldsymbol{\beta}}(\cdot)$ denotes the p.d.f of $\boldsymbol{x}^T\boldsymbol{\beta}$. Note that

$$\begin{aligned}
\mathrm{E}\gamma_i(z_j)=&\mathrm{E}\Big\{K\Big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}-\boldsymbol{x}_j^T\boldsymbol{\beta}}{h}\Big)\big[f_0(\boldsymbol{x}_j^T\boldsymbol{\beta}_0)-G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]\Big\}\\
=&\mathrm{E}_{(\boldsymbol{x}_i^T\boldsymbol{\beta},\boldsymbol{x}_j^T\boldsymbol{\beta})}\Big\{\mathrm{E}\Big\{K\Big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}-\boldsymbol{x}_j^T\boldsymbol{\beta}}{h}\Big)\big[f_0(\boldsymbol{x}_j^T\boldsymbol{\beta}_0)-G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]\Big|\boldsymbol{x}_i^T\boldsymbol{\beta},\boldsymbol{x}_j^T\boldsymbol{\beta}\Big\}\Big\}\\
=&\mathrm{E}_{(\boldsymbol{x}_i^T\boldsymbol{\beta},\boldsymbol{x}_j^T\boldsymbol{\beta})}\Big\{K\Big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}-\boldsymbol{x}_j^T\boldsymbol{\beta}}{h}\Big)\big[G(\boldsymbol{x}_j^T\boldsymbol{\beta}|\boldsymbol{\beta})-G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]\Big\}\\
=&\mathrm{E}_{\boldsymbol{x}_i^T\boldsymbol{\beta}}\Big\{h\int K(-z)\big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}+hz|\boldsymbol{\beta})-G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]f_{\boldsymbol{\beta}}(\boldsymbol{x}_i^T\boldsymbol{\beta}+hz)dz\Big\}\\
=&\mathrm{E}_{\boldsymbol{x}_i^T\boldsymbol{\beta}}\Big\{h\int K(-z)\Big[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})hz+\frac{h^2z^2}{2}G^{(2)}(t_1|\boldsymbol{\beta})\Big]\Big[f_{\boldsymbol{\beta}}(\boldsymbol{x}_i^T\boldsymbol{\beta})+hzf'_{\boldsymbol{\beta}}(\widetilde{t})\Big]dz\Big\}\\
=&\mathrm{E}_{\boldsymbol{x}_i^T\boldsymbol{\beta}}\Big\{\frac{h^3f_{\boldsymbol{\beta}}(\boldsymbol{x}_i^T\boldsymbol{\beta})}{2}\int z^2K(-z)G^{(2)}(t_1|\boldsymbol{\beta})dz\Big\}+\mathrm{E}_{\boldsymbol{x}_i^T\boldsymbol{\beta}}\Big\{h^3G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\int z^2K(-z)f'_{\boldsymbol{\beta}}(\widetilde{t})dz\Big\}\\
&+\mathrm{E}_{\boldsymbol{x}_i^T\boldsymbol{\beta}}\Big\{\frac{h^4}{2}\int z^3K(-z)G^{(2)}(t_1|\boldsymbol{\beta})f'_{\boldsymbol{\beta}}(\widetilde{t})dz\Big\},
\end{aligned}$$

where t_1 and $\widetilde{t}$ are both between $\boldsymbol{x}_i^T\boldsymbol{\beta}$ and $\boldsymbol{x}_i^T\boldsymbol{\beta}+hz$. In the above, the third equality applies the independence between $\boldsymbol{x}_i^T\boldsymbol{\beta}$ and $\boldsymbol{x}_j^T\boldsymbol{\beta}$, and $G(\boldsymbol{x}_j^T\boldsymbol{\beta}|\boldsymbol{\beta})=\mathrm{E}[f_0(\boldsymbol{x}_j^T\boldsymbol{\beta}_0)|\boldsymbol{x}_j^T\boldsymbol{\beta}]$. According to (A3)–(A5), we know that $\sup_{\boldsymbol{\beta}\in\mathbb{B}}\mathrm{E}\big[A_{n2}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]\leqslant ch^2$ for some c large enough.

Observing that $\sup_{\boldsymbol{\beta}\in\mathbb{B}}\mathrm{E}[\gamma_i(z_j)^2]\leqslant c_1h$, for some positive constant c_1. Since $K(\cdot)$ is bounded on the real line, for any fixed $\boldsymbol{\beta}$, by Bernstein's inequality, $\forall\, t\geqslant 0$, there exists some constant $c>0$ such that

$$P\left(\left|\sum_{j=1,j\neq i}^n\gamma_i(z_j)-\mathrm{E}\gamma_i(z_j)\right|\geqslant t\mid\boldsymbol{x}_i^T\boldsymbol{\beta}\right)\leqslant 2\exp\left[\frac{-ct^2}{(n-1)h}\right].$$

Note that $\mathrm{E}\gamma(z_i) \leqslant ch^3$ for some positive constant c. Take $t = (n-1)h^3$, then we have that

$$
\begin{aligned}
& P\left(\left|\sum_{j=1,j\neq i}^{n} \gamma_i(z_j) - \mathrm{E}\gamma_i(z_j)\right| \geqslant (n-1)h^3\right) \\
=& \mathrm{E}_{\boldsymbol{x}_i^T\boldsymbol{\beta}}\left\{P\left(\left|\sum_{j=1,j\neq i}^{n} \gamma_i(z_j) - \mathrm{E}\gamma_i(z_j)\right| \geqslant (n-1)h^3 \mid \boldsymbol{x}_i^T\boldsymbol{\beta}\right)\right\} \\
\leqslant& 2\exp\left[-c(n-1)h^5\right].
\end{aligned}
$$

Hence we can conclude that $P\Big(\big|A_{n2}(t|\boldsymbol{\beta})\big| \geqslant c_0h^2\Big) \leqslant 2\exp(-c_1nh^5)$ for some positive constants c_0, c_1, and all n sufficiently large.

To obtain the uniform bound, we cover $\mathbb{B}$ with L_2-balls with radius δ^2. Let $\mathcal{N}_{\delta^2}$ be the δ^2-cover of $\mathbb{B}$. The covering number $N_2 = |\mathcal{N}_{\delta^2}|$ satisfies $N_2 \leqslant cp^{2ks}\delta^{-4ks}$ for sufficiently large c, as shown in the proof of Lemma B4. For any $\boldsymbol{\beta}$ in such a ball with center $\boldsymbol{\beta}^*$, we need to bound

$$
\begin{aligned}
& \Bigg|(n-1)^{-1}\sum_{j=1,j\neq i}^{n} K_h(\boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{x}_j^T\boldsymbol{\beta})\big[f_0(\boldsymbol{x}_j^T\boldsymbol{\beta}_0) - G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big] \\
& - (n-1)^{-1}\sum_{j=1,j\neq i}^{n} K_h(\boldsymbol{x}_i^T\boldsymbol{\beta}^* - \boldsymbol{x}_j^T\boldsymbol{\beta}^*)\big[f_0(\boldsymbol{x}_j^T\boldsymbol{\beta}_0) - G(\boldsymbol{x}_i^T\boldsymbol{\beta}^*|\boldsymbol{\beta}^*)\big]\Bigg| \\
\leqslant& \Bigg|(n-1)^{-1}\sum_{j=1,j\neq i}^{n}\Big[K_h(\boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{x}_j^T\boldsymbol{\beta}) - K_h(\boldsymbol{x}_i^T\boldsymbol{\beta}^* - \boldsymbol{x}_j^T\boldsymbol{\beta}^*)\Big]\big[G(\boldsymbol{x}_j^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0) - G(\boldsymbol{x}_j^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]\Bigg| \\
& + \Bigg|(n-1)^{-1}\sum_{j=1,j\neq i}^{n}\Big[K_h(\boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{x}_j^T\boldsymbol{\beta}) - K_h(\boldsymbol{x}_i^T\boldsymbol{\beta}^* - \boldsymbol{x}_j^T\boldsymbol{\beta}^*)\Big]\big[G(\boldsymbol{x}_j^T\boldsymbol{\beta}|\boldsymbol{\beta}) - G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\big]\Bigg| \\
& + \Bigg|(n-1)^{-1}\sum_{j=1,j\neq i}^{n} K_h(\boldsymbol{x}_i^T\boldsymbol{\beta}^* - \boldsymbol{x}_j^T\boldsymbol{\beta}^*)\big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - G(\boldsymbol{x}_i^T\boldsymbol{\beta}^*|\boldsymbol{\beta}^*)\big]\Bigg| \\
=& \sum_{k=1}^{3}|I_{nk}|,
\end{aligned}
$$

where the definition of I_{nk} is clear from the context.

Lemma B2 implies that

$$P\left(\max_{1\leqslant i\leqslant n}|\boldsymbol{x}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}^*)|\geqslant\sigma_x\sqrt{s\log(p\vee n)}||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2,\ \forall\boldsymbol{\beta},\boldsymbol{\beta}^*\in\mathbb{B}\right)\leqslant\exp[-cs\log(p\vee n)],$$

for some positive constant c, and all n sufficiently large. Lemma B2 implies that

$$\begin{aligned}P\Bigg(&\max_{1\leqslant i\leqslant n}\left|(n-1)^{-1}\sum_{j=1,j\neq i}^{n}|\boldsymbol{x}_j^T(\boldsymbol{\beta}-\boldsymbol{\beta}^*)|^2-(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^T\mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T)(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\right|\\&\qquad\qquad\geqslant c_0\sigma_x^2\sqrt{\frac{s\log(p\vee n)}{n}}||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2^2,\forall\ \boldsymbol{\beta},\boldsymbol{\beta}^*\in\mathbb{B}\Bigg)\\&\leqslant\exp[-cs\log(p\vee n)],\end{aligned}$$

for some positive constants c_0, c, and all n sufficiently large. Assumption (A2) implies that $(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^T\mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T)(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\leqslant\xi_3||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2^2$. Hence we have that

$$\begin{aligned}\max_{1\leqslant i\leqslant n}(n-1)^{-1}\sum_{j=1,j\neq i}^{n}|\boldsymbol{x}_j^T(\boldsymbol{\beta}-\boldsymbol{\beta}^*)|^2&\leqslant\left(\xi_3+c_0\sigma_x^2\sqrt{\frac{s\log(p\vee n)}{n}}\right)||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2^2\\&\leqslant\sigma_x^2 s\log(p\vee n)||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2^2,\ \forall\ \boldsymbol{\beta},\boldsymbol{\beta}^*\in\mathbb{B}\end{aligned}$$

with probability at least $1-\exp[-c\log(p\vee n)]$, for some positive constants c_0, c, and all n sufficiently large.

Denote the event

$$\begin{aligned}\mathcal{E}_2=&\Big\{\max_{1\leqslant i\leqslant n}|\boldsymbol{x}_i^T\boldsymbol{\beta}|^2+(n-1)^{-1}\sum_{j=1,j\neq i}^{n}|\boldsymbol{x}_j^T\boldsymbol{\beta}|^2\leqslant 2\sigma_x^2 s\log(p\vee n)||\boldsymbol{\beta}||_2^2,\ \forall\ \boldsymbol{\beta}\in\mathbb{B}\Big\}\\&\bigcap\Big\{\max_{1\leqslant i\leqslant n}|\boldsymbol{x}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}^*)|^2+(n-1)^{-1}\sum_{j=1,j\neq i}^{n}|\boldsymbol{x}_j^T(\boldsymbol{\beta}-\boldsymbol{\beta}^*)|^2\leqslant 2\sigma_x^2 s\log(p\vee n)||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2^2,\ \forall\ \boldsymbol{\beta},\boldsymbol{\beta}^*\in\mathbb{B}\Big\}\\&\bigcap\Big\{\max_{1\leqslant i\leqslant n}\left|G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-G(\boldsymbol{x}_i^T\boldsymbol{\beta}^*|\boldsymbol{\beta}^*)\right|\leqslant c_0\sqrt{s\log(p\vee n)}||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2^{1/2},\ \forall\ \boldsymbol{\beta},\boldsymbol{\beta}^*\in\mathbb{B}\Big\}\\&\bigcap\Big\{n^{-1}\sum_{i=1}^{n}\left|G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right|^2\leqslant c_0||\boldsymbol{\beta}-\boldsymbol{\beta}_0||_2^2,\ \forall\ \boldsymbol{\beta}\in\mathbb{B}\Big\}.\end{aligned}$$

Hence $P(\mathcal{E}_2) \geqslant 1-6\exp[-c_1\log(p\vee n)]$, for some positive constant c_0, c_1 and all n sufficiently large. In the above, the third event applies (S4) in Lemma A4, and the fourth event applies (S6) in Lemma A4.

Take $\delta = \frac{h^3}{4\sqrt{s\log(p\vee n)}}$. There exist positive constants c_0, c_1, such that for all n sufficiently large,

$$
\begin{aligned}
|I_{n1}| \leqslant & c_1[(n-1)h^2]^{-1}\sum_{j=1,j\neq i}^{n}|(\boldsymbol{x}_i-\boldsymbol{x}_j)^T(\boldsymbol{\beta}-\boldsymbol{\beta}^*)| * |G(\boldsymbol{x}_j^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)-G(\boldsymbol{x}_j^T\boldsymbol{\beta}|\boldsymbol{\beta})| \\
\leqslant & c_1h^{-2}\sqrt{|\boldsymbol{x}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}^*)|^2+(n-1)^{-1}\sum_{j=1,j\neq i}^{n}|\boldsymbol{x}_j^T(\boldsymbol{\beta}-\boldsymbol{\beta}^*)|^2} \\
& * \sqrt{(n-1)^{-1}\sum_{j=1,j\neq i}^{n}|G(\boldsymbol{x}_j^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)-G(\boldsymbol{x}_j^T\boldsymbol{\beta}|\boldsymbol{\beta})|^2} \\
\leqslant & c_1h^{-2}||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2 * ||\boldsymbol{\beta}-\boldsymbol{\beta}_0||_2\sqrt{s\log(p\vee n)} \\
\leqslant & c_0h^{-2}\delta^2\sqrt{s\log(p\vee n)} = \frac{c_0h^4}{16\sqrt{s\log(p\vee n)}}.
\end{aligned}
$$

on the event $\mathcal{E}_2$. In the above, the second last inequality applies the second and the fourth events in $\mathcal{E}_2$.

For I_{n2}, the Lipschitz condition for $K(\cdot)$ implies that there exists $\widetilde{t}_j$ between $\boldsymbol{x}_i^T\boldsymbol{\beta}$ and $\boldsymbol{x}_j^T\boldsymbol{\beta}$ such that on the event $\mathcal{E}_2$,

$$
\begin{aligned}
|I_{n2}| &\leqslant c_1[(n-1)h^2]^{-1}\sum_{j=1,j\neq i}^{n}|(\boldsymbol{x}_i-\boldsymbol{x}_j)^T(\boldsymbol{\beta}-\boldsymbol{\beta}^*)| * \left|G^{(1)}(\widetilde{t}_j|\boldsymbol{\beta})(\boldsymbol{x}_i-\boldsymbol{x}_j)^T\boldsymbol{\beta}\right| \\
&\leqslant c_1h^{-2}||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2 * ||\boldsymbol{\beta}||_2 s\log(p\vee n) \\
&\leqslant c_0h^{-2}\delta^2 s\log(p\vee n) = \frac{c_0h^4}{16},
\end{aligned}
$$

for all n sufficiently large. In the above, the second inequality applies the first and the second events in $\mathcal{E}_2$. Assumption (A5)-(a) implies that $\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}}|G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta})| \leqslant b$ and $G^{(2)}(t|\boldsymbol{\beta})$ is bounded for any $t\in\mathbb{R}$, $\boldsymbol{\beta}\in\mathbb{B}$. It indicates that $\max_{i\neq j}\sup_{\boldsymbol{\beta}\in\mathbb{B}}|G^{(1)}(\widetilde{t}_j|\boldsymbol{\beta})| \leqslant c$

for positive constant c.

For $|I_{n3}|$, we have

$$
\begin{aligned}
|I_{n3}| &\leqslant c_1 h^{-1} |G(\boldsymbol{x}_i^T \boldsymbol{\beta} | \boldsymbol{\beta}) - G(\boldsymbol{x}_i^T \boldsymbol{\beta}^* | \boldsymbol{\beta}^*)| \\
&\leqslant c_1 h^{-1} ||\boldsymbol{\beta} - \boldsymbol{\beta}^*||_2^{1/2} \sqrt{s \log(p \vee n)} \\
&\leqslant c_0 h^{-1} \delta \sqrt{s \log(p \vee n)} = \frac{c_0 h^2}{4},
\end{aligned}
$$

on the event $\mathcal{E}_2$, for positive constants c_0, c_1, and all n sufficiently large. In the above, the second inequality applies the third event in $\mathcal{E}_2$. Combining all the previous results, we conclude that

$$
\begin{aligned}
&\Bigg|(n-1)^{-1} \sum_{j=1, j \neq i}^{n} K_h(\boldsymbol{x}_i^T \boldsymbol{\beta} - \boldsymbol{x}_j^T \boldsymbol{\beta}) \big[f_0(\boldsymbol{x}_j^T \boldsymbol{\beta}_0) - G(\boldsymbol{x}_i^T \boldsymbol{\beta} | \boldsymbol{\beta}) \big] \\
&- (n-1)^{-1} \sum_{j=1, j \neq i}^{n} K_h(\boldsymbol{x}_i^T \boldsymbol{\beta}^* - \boldsymbol{x}_j^T \boldsymbol{\beta}^*) \big[f_0(\boldsymbol{x}_j^T \boldsymbol{\beta}_0) - G(\boldsymbol{x}_i^T \boldsymbol{\beta}^* | \boldsymbol{\beta}^*) \big] \Bigg| \leqslant c_0 h^2/2,
\end{aligned}
$$

on the event $\mathcal{E}_2$, for all n sufficiently large. Then it implies that

$$
\begin{aligned}
& P\left(\sup_{\boldsymbol{\beta} \in \mathbb{B}} |A_{n2}(\boldsymbol{x}_i^T \boldsymbol{\beta} | \boldsymbol{\beta})| \geqslant c_0 h^2 \right) \\
\leqslant & P\left(\bigcup_{\boldsymbol{\beta}^* \in \mathcal{N}_{\delta 2}} \big| A_{n2}(\boldsymbol{x}_i^T \boldsymbol{\beta}^* | \boldsymbol{\beta}^*) \big| \geqslant c_0 h^2/2 \right) \\
& + P\left(\sup_{\boldsymbol{\beta}^* \in \mathcal{N}_{\delta 2}} \sup_{||\boldsymbol{\beta} - \boldsymbol{\beta}^*||_2 \leqslant \delta^2} |A_{n2}(\boldsymbol{x}_i^T \boldsymbol{\beta} | \boldsymbol{\beta}) - A_{n2}(\boldsymbol{x}_i^T \boldsymbol{\beta}^* | \boldsymbol{\beta}^*)| \leqslant c_0 h^2/2 \right) \\
\leqslant & \sum_{\boldsymbol{\beta}^* \in \mathcal{N}_{\delta 2}} P\left(\big| A_{n2}(\boldsymbol{x}_i^T \boldsymbol{\beta}^* | \boldsymbol{\beta}^*) \big| \geqslant c_0 h^2/2 \right) + 5 \exp[-c \log(p \vee n)] \\
\leqslant & c p^{2ks} \delta^{-4ks} \exp(-c n h^5) + 5 \exp[-c \log(p \vee n)] \\
\leqslant & \exp[-c_2 \log(p \vee n)],
\end{aligned}
$$

for some positive constants c, c_0, and $c_2 > 1$, and all n sufficiently large. We thus conclude

$$\begin{aligned} P\Big(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}}|A_{n2}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})|\geqslant c_0h^2\Big) \leqslant & \sum_{i=1}^n P\Big(\sup_{\boldsymbol{\beta}\in\mathbb{B}}|A_{n2}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})|\geqslant c_0h^2\Big) \\ \leqslant & n\exp[-c_2\log(p\vee n)] = \exp\{-[c_2\log(p\vee n)-\log n]\} \\ \leqslant & \exp[-c_1\log(p\vee n)], \end{aligned}$$

for positive constants c_0 c_1, and all n sufficiently large. □

Lemma B6. *Under the assumptions of Theorem 1, there exist some positive constants c_0, c_1 such that for all n sufficiently large,*

$$P\left(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}}\left|A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-\mathrm{E}\left[A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\right]\right|\geqslant c_0h^2\right)\leqslant\exp(-c_1nh^5).$$

Proof. Let $A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) = [(n-1)h]^{-1}\sum_{j=1,j\neq i}^n z_{ij}(\boldsymbol{\beta})$, where $z_{ij}(\boldsymbol{\beta}) = K\Big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}-\boldsymbol{x}_j^T\boldsymbol{\beta}}{h}\Big)$. Observing that $\sup_{\boldsymbol{\beta}\in\mathbb{B}}\mathrm{E}[z_{ij}^2(\boldsymbol{\beta})]\leqslant c_1h$. Since $K(\cdot)$ is bounded on the real line, then for any fixed $\boldsymbol{\beta}$, by Bernstein's inequality, $\forall\ \eta>0$, there exists constant $c>0$ such that

$$P\left(\left|\sum_{j=1,j\neq i}^n z_{ij}(\boldsymbol{\beta})-\mathrm{E}z_{ij}(\boldsymbol{\beta})\right|\geqslant\eta\right)\leqslant 2\exp\left[\frac{-c\eta^2}{(n-1)h+\eta}\right].$$

Take $\eta=(n-1)h^3$, then we can conclude that $P\left(\left|A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-\mathrm{E}\left[A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\right]\right|\geqslant c_0h^2\right)\leqslant 2\exp(-c_1nh^5)$ for positive constants c_0 and c_1.

To obtain the uniform bound, we cover $\mathbb{B}$ with L_2-balls with radius δ. Let $\mathcal{N}_\delta$ be an $\delta-$cover of $\mathbb{B}$. The covering number $N=|\mathcal{N}_\delta|$ satisfies $N\leqslant cp^{2ks}\delta^{-2ks}$ for sufficiently large c, as shown in the proof of Lemma B4. For any $\boldsymbol{\beta}$ in such a ball with center $\boldsymbol{\beta}^*$, let us take

$\delta = \frac{h^4}{4\sqrt{n}}$, then the Lipschitz condition for $K(\cdot)$ implies that

$$\begin{aligned}
&\left|A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}^*|\boldsymbol{\beta}^*)\right| \\
=&\left|(n-1)^{-1}\sum_{j=1,j\neq i}^{n}\left[K_h(\boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{x}_j^T\boldsymbol{\beta}) - K_h(\boldsymbol{x}_i^T\boldsymbol{\beta}^* - \boldsymbol{x}_j^T\boldsymbol{\beta}^*)\right]\right| \\
\leqslant&[c(n-1)h^2]^{-1}\sum_{j=1,j\neq i}^{n}\left|(\boldsymbol{x}_i - \boldsymbol{x}_j)^T(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\right| \leqslant c_0h^2/4,
\end{aligned}$$

with probability at least $1-\exp(-cn)$, for some positive constants c, c_0, and all n sufficiently large, similarly as the proof of Lemma B4. It also implies that

$$\begin{aligned}
&\left|\mathrm{E}\left[A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\right] - \mathrm{E}\left[A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}^*|\boldsymbol{\beta}^*)\right]\right| \\
=&\left|\mathrm{E}_{(\boldsymbol{x}_i^T\boldsymbol{\beta},\boldsymbol{x}_i^T\boldsymbol{\beta}^*)}\left[A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}^*|\boldsymbol{\beta}^*)\right]\right| \\
\leqslant&c_0h^2/4,
\end{aligned}$$

with probability at least $1-\exp(-cn)$, for some positive constants c, c_0, and all n sufficiently large. Then it implies that

$$\begin{aligned}
&P\left(\sup_{\boldsymbol{\beta}\in\mathbb{B}}\left|A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - \mathrm{E}\left[A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\right]\right| \geqslant c_0h^2\right) \\
\leqslant&P\left(\bigcup_{\boldsymbol{\beta}^*\in\mathcal{N}_\delta}\left|A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}^*|\boldsymbol{\beta}^*) - \mathrm{E}\left[A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}^*|\boldsymbol{\beta}^*)\right]\right| \geqslant c_0h^2/2\right) \\
&+P\left(\sup_{\boldsymbol{\beta}^*\in\mathcal{N}_{\delta^2}}\sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2\leqslant\delta^2}\left|A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}^*|\boldsymbol{\beta}^*)\right| \leqslant c_0h^2/4\right) \\
&+P\left(\sup_{\boldsymbol{\beta}^*\in\mathcal{N}_{\delta^2}}\sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2\leqslant\delta^2}\left|\mathrm{E}\left[A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\right] - \mathrm{E}\left[A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}^*|\boldsymbol{\beta}^*)\right]\right| \leqslant c_0h^2/4\right) \\
\leqslant&\sum_{\boldsymbol{\beta}^*\in\mathcal{N}_\delta}P\left(\left|A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}^*|\boldsymbol{\beta}^*) - \mathrm{E}\left[A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}^*|\boldsymbol{\beta}^*)\right]\right| \geqslant c_0h^2/2\right) + \exp(-cn) \\
\leqslant&cp^{2ks}\delta^{-2ks}\exp(-cnh^5) + \exp(-cn) \leqslant \exp(-c_1nh^5).
\end{aligned}$$

for some positive constants c, c_0, c_1, and all n sufficiently large. We conclude

$$\begin{aligned}
&P\left(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}}\left|A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-\mathrm{E}\left[A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\right]\right|\geqslant c_0h^2\right)\\
\leqslant&\sum_{i=1}^n P\left(\sup_{\boldsymbol{\beta}\in\mathbb{B}}\left|A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-\mathrm{E}\left[A_{n3}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})\right]\right|\geqslant c_0h^2\right)\\
\leqslant& n\exp(-c_1nh^5)=\exp[-(c_1nh^5-\log n)]\\
\leqslant&\exp(-c_2nh^5),
\end{aligned}$$

for positive constants c_0, c_1, c_2, and all n sufficiently large. □

Lemma B7. *Under the assumptions of Theorem 1 and Lemma 2, there exist universal positive constants c_0, c_1 such that for all n sufficiently large,*

$$P\left\{\sup_{\substack{\boldsymbol{\beta}\in\mathbb{B}\\ \boldsymbol{v}\in\mathbb{K}(p-1,2ks+\widetilde{s})}}\left(n^{-1}\sum_{i=1}^n\boldsymbol{v}^T\left[\boldsymbol{x}_{i,-1}-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})\right]\left[\boldsymbol{x}_{i,-1}-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})\right]^T\boldsymbol{v}\right)\geqslant c_0\right\}\leqslant\exp(-c_1n),$$

where $\mathbb{K}(p-1,2ks+\widetilde{s})=\{\boldsymbol{v}\in\mathbb{R}^{p-1}:||\boldsymbol{v}||_2\leqslant 1,||\boldsymbol{v}||_0\leqslant 2ks+\widetilde{s}\}$*, and* $\widetilde{s}=\max_{2\leqslant j\leqslant p}||\boldsymbol{d}_{0j}||_0$.

Proof. We observe that

$$\begin{aligned}
&\sup_{\substack{\boldsymbol{\beta}\in\mathbb{B}\\ \boldsymbol{v}\in\mathbb{K}(p-1,2ks+\widetilde{s})}} n^{-1}\sum_{i=1}^n\boldsymbol{v}^T\Big[\boldsymbol{x}_{i,-1}-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})\Big]\Big[\boldsymbol{x}_{i,-1}-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})\Big]^T\boldsymbol{v}\\
\leqslant&\sup_{\boldsymbol{v}\in\mathbb{K}(p-1,2ks+\widetilde{s})}2n^{-1}\sum_{i=1}^n\boldsymbol{v}^T\Big[\boldsymbol{x}_{i,-1}-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\Big]\Big[\boldsymbol{x}_{i,-1}-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\Big]^T\boldsymbol{v}\\
&+\sup_{\substack{\boldsymbol{\beta}\in\mathbb{B}\\ \boldsymbol{v}\in\mathbb{K}(p-1,2ks+\widetilde{s})}}2n^{-1}\sum_{i=1}^n\boldsymbol{v}^T\Big[\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})\Big]\Big[\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta})\Big]^T\boldsymbol{v}\\
\triangleq&2J_{n1}+2J_{n2},
\end{aligned}$$

where the definition for J_{nk} is clear from the context.

Note that $\boldsymbol{x}_{i,-1}-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)$ is sub-Gaussian with covariance matrix $\mathrm{E}\big[\mathrm{Cov}(\boldsymbol{x}_{-1}|\boldsymbol{x}^T\boldsymbol{\beta}_0)\big]$.

Assumption (A2) implies that

$$J_{n1} \leqslant \xi_1 + \sup_{\boldsymbol{v} \in \mathbb{K}(p-1, 2ks+\widetilde{s})} \left| \boldsymbol{v}^T \boldsymbol{\Psi}_n \boldsymbol{v} \right|,$$

where $\boldsymbol{\Psi}_n = n^{-1} \sum_{i=1}^n \left[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\right]\left[\boldsymbol{x}_{i,-1} - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\right]^T - \mathrm{E}\left[\mathrm{Cov}(\boldsymbol{x}_{-1}|\boldsymbol{x}^T\boldsymbol{\beta}_0)\right]$. Lemma B2 implies that

$$P\left(\sup_{\boldsymbol{v} \in \mathbb{K}(p-1, 2ks+\widetilde{s})} \left| \boldsymbol{v}^T \boldsymbol{\Psi}_n \boldsymbol{v} \right| \geqslant \sigma_x^2\right) \leqslant \exp(-c_1 n),$$

for some universal positive constant c_1, and all n sufficiently large, since $s \log p = o(n)$, and $\widetilde{s} \log p = o(n)$. Hence with probability at least $1 - \exp(-c_1 n)$, we have $J_{n1} \leqslant \xi_1 + \sigma_x^2$.

To bound J_{n2}, Assumption (A2)-(c) implies that

$$\begin{aligned}
J_{n2} &\leqslant \sup_{\boldsymbol{\beta} \in \mathbb{B}} c^2 n^{-1} \sum_{i=1}^n \left[|\boldsymbol{x}_i^T\boldsymbol{\beta}_0 - \boldsymbol{x}_i^T\boldsymbol{\beta}| + (|\boldsymbol{x}_i^T\boldsymbol{\beta}| + |\boldsymbol{x}_i^T\boldsymbol{\beta}_0|)||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_2\right]^2 \\
&\leqslant \sup_{\boldsymbol{\beta} \in \mathbb{B}} \frac{2c^2}{n} \sum_{i=1}^n (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \boldsymbol{x}_i \boldsymbol{x}_i^T (\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \sup_{\boldsymbol{\beta} \in \mathbb{B}} \frac{4c^2}{n} ||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_2^2 \sum_{i=1}^n \left[(\boldsymbol{x}_i^T\boldsymbol{\beta})^2 + (\boldsymbol{x}_i^T\boldsymbol{\beta}_0)^2\right],
\end{aligned}$$

for some positive constant c. Since $\boldsymbol{\beta} \in \mathbb{B}$, we have $||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_2 \leqslant r$. Note that $\boldsymbol{x}_i$ is also sub-Gaussian. Combining Lemma B2 and similar technique as above, we have

$$P\left\{n^{-1} \sum_{i=1}^n [\boldsymbol{x}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0)]^2 \leqslant (\xi_3 + \sigma_x^2)||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_2,\ \forall\ \boldsymbol{\beta} \in \mathbb{B}\right\} \geqslant 1 - \exp(-c_1 n),$$

$$P\left\{n^{-1} \sum_{i=1}^n |\boldsymbol{x}_i^T\boldsymbol{\beta}|^2 \leqslant (\xi_3 + \sigma_x^2)||\boldsymbol{\beta}||_2,\ \forall\ \boldsymbol{\beta} \in \mathbb{B}\right\} \geqslant 1 - \exp(-c_1 n),$$

for some positive constant c_1 and all n sufficiently large. We thus have that $J_{n2} \leqslant c_0 r^2$, with probability at least $1 - 2\exp(-c_1 n)$, for some positive constants c_0, c_1, and all n sufficiently large. Since $r \leqslant 1$, the conclusion follows. $\square$

Lemma B8. *Under assumptions (A1) and (A5), if* $\log p = O(n)$, *then there exist some* *positive constants c_1, c_2, such that for all n sufficiently large,*

$$P\left(\max_{2\leqslant j\leqslant p}\left\|\frac{1}{n}\sum_{i=1}^{n}[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]^2\widetilde{\boldsymbol{x}}_{i,-j*}\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\phi}_{0j}\right\|_\infty \geqslant c_1\sigma_x^2\sqrt{\frac{\log p}{n}}\right)\leqslant \exp(-c_2\log p),$$

where $\boldsymbol{\phi}_{0j}=\tau_{0j}^2\boldsymbol{\theta}_j$.

Proof. Note that for any j, $G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\phi}_{0j}$ and $G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\widetilde{\boldsymbol{x}}_{i,-j*}$ are both sub-Gaussian with the variance proxy no larger than $2b^4\xi_1^2\xi_2^{-2}\sigma_x^2$ and $2b^2\sigma_x^2$, respectively, by Lemma A2 and Lemma A13. The definition of $\boldsymbol{d}_{0j}$ implies that $\mathrm{E}\{[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]^2\widetilde{\boldsymbol{x}}_{i,-j*}\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\phi}_{0j}\}=\mathbf{0}_{p-2}$. Hence Lemma B1 implies that

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]^2\widetilde{\boldsymbol{x}}_{i,-j*}\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\phi}_{0j}\right|\geqslant 2c_0b^3\xi_1\xi_2^{-1}\sigma_x^2\sqrt{\frac{\log p}{n}}\right)\leqslant \exp(-c_1\log p),$$

for some positive constants c_0, $c_1>1$, and all n sufficiently large. Note that $2b^3\xi_1\xi_2^{-1}$ is a positive constant that does not depend on $\boldsymbol{x}_i$. Then we have

$$\begin{aligned}&P\left(\max_{2\leqslant j\leqslant p}\left\|\frac{1}{n}\sum_{i=1}^{n}[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]^2\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\phi}_{0j}\widetilde{\boldsymbol{x}}_{i,-j*}\right\|_\infty\geqslant c_1\sigma_x^2\sqrt{\frac{\log p}{n}}\right)\\
&\leqslant\sum_{j=2}^{p}P\left(\left|\frac{1}{n}\sum_{i=1}^{n}[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]^2\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\phi}_{0j}\widetilde{\boldsymbol{x}}_{i,-j*}\right|\geqslant c_1\sigma_x^2\sqrt{\frac{\log p}{n}}\right)\\
&\leqslant\exp(-c_2\log p),\end{aligned}$$

for some positive constants c_1, c_2, and all n sufficiently large. □

Lemma B9. *Assume the conditions of Lemma 2 are satisfied, then there exist universal positive constants c_0 and c_1 such that for all n sufficiently large,*

$$P\left(\max_{2\leqslant j\leqslant p}\left\|\frac{1}{n}\sum_{i=1}^{n}\{[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2-[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]^2\}\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\phi}_{0j}\right\|_\infty\geqslant c_0h\right)\leqslant\exp(-c_1\log p),$$

$$P\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\{[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2-[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]^2\}\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}\right\|_\infty\geqslant c_0h\right)\leqslant\exp(-c_1\log p),$$

where $\boldsymbol{\phi}_{0j} = \tau_{0j}^2 \boldsymbol{\theta}_j$.

Proof. We will prove the first part of the claim below. The proof of the second part is similar. The Cauchy-Schwartz inequality implies that

$$
\begin{aligned}
&\Big\| \frac{1}{n} \sum_{i=1}^{n} \{ [\widehat{G}^{(1)}(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}} | \widehat{\boldsymbol{\beta}})]^2 - [G^{(1)}(\boldsymbol{x}_i^T \boldsymbol{\beta}_0 | \boldsymbol{\beta}_0)]^2 \} \widetilde{\boldsymbol{x}}_{i,-1} \widetilde{\boldsymbol{x}}_{i,-1}^T \boldsymbol{\phi}_{0j} \Big\|_\infty \\
\leqslant & \max_{1 \leqslant i \leqslant n} \Big| [\widehat{G}^{(1)}(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}} | \widehat{\boldsymbol{\beta}})]^2 - [G^{(1)}(\boldsymbol{x}_i^T \boldsymbol{\beta}_0 | \boldsymbol{\beta}_0)]^2 \Big| * \sqrt{\max_{2 \leqslant j \leqslant p} n^{-1} \sum_{i=1}^{n} \widetilde{x}_{i,j}^2} \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\widetilde{\boldsymbol{x}}_{i,-1}^T \boldsymbol{\phi}_{0j})^2}.
\end{aligned}
$$

Lemma A11 and Assumption (A5)-(a) together imply that

$$
\begin{aligned}
&P\left(\max_{1 \leqslant i \leqslant n} \big| [\widehat{G}^{(1)}(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}} | \widehat{\boldsymbol{\beta}})]^2 - [G^{(1)}(\boldsymbol{x}_i^T \boldsymbol{\beta}_0 | \boldsymbol{\beta}_0)]^2 \big| \geqslant ch \right) \\
\leqslant & P\left(\max_{1 \leqslant i \leqslant n} \big| \widehat{G}^{(1)}(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}} | \widehat{\boldsymbol{\beta}}) - G^{(1)}(\boldsymbol{x}_i^T \boldsymbol{\beta}_0 | \boldsymbol{\beta}_0) \big| \geqslant c_0 h \right) \\
&+ P\left(\max_{1 \leqslant i \leqslant n} \big| \widehat{G}^{(1)}(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}} | \widehat{\boldsymbol{\beta}}) + G^{(1)}(\boldsymbol{x}_i^T \boldsymbol{\beta}_0 | \boldsymbol{\beta}_0) \big| \geqslant 2b + c_0 h \right) \\
\leqslant & 2 \exp(-d_1 \log p),
\end{aligned}
\tag{S50}
$$

for some universal positive constants c, c_0, d_1, and all n sufficiently large. Note that Lemma A2 implies that $\widetilde{x}_{i,j}$ is sub-Gaussian with variance proxy at most $2\sigma_x^2$. Hence $\mathrm{E}(\widetilde{x}_{i,j}^2) \leqslant 2\sigma_x^2$ uniformly in j. Lemma B1 implies that there exist some positive constants d_1, d_2, such that for all n sufficiently large,

$$
\begin{aligned}
P\left(\max_{2 \leqslant j \leqslant p} n^{-1} \sum_{i=1}^{n} \widetilde{x}_{i,j}^2 \geqslant 3\sigma_x^2 \right) &\leqslant P\left(\max_{2 \leqslant j \leqslant p} n^{-1} \sum_{i=1}^{n} \widetilde{x}_{i,j}^2 \geqslant \max_{2 \leqslant j \leqslant p} \mathrm{E}(\widetilde{x}_{i,j}^2) + d_1 \sigma_x^2 \sqrt{\frac{\log p}{n}} \right) \\
&\leqslant \exp(-d_2 \log p).
\end{aligned}
$$

Since $\widetilde{\boldsymbol{x}}_{i,-1}^T \boldsymbol{\phi}_{0j}$ is sub-Gaussian with variance proxy at most $C\sigma_x^2$ for some constant $C > 0$, Lemma B1 also implies that $P\left(\frac{1}{n} \sum_{i=1}^{n} (\widetilde{\boldsymbol{x}}_{i,-1}^T \boldsymbol{\phi}_{0j})^2 \geqslant 2\xi_1 \xi_2^{-2} \right) \leqslant \exp(-d_2 \log p)$, for some positive constants d_1, d_2, and all n sufficiently large.

Hence we have

$$
\begin{aligned}
&P\left(\max_{2\leqslant j\leqslant p}\Big\|\frac{1}{n}\sum_{i=1}^n\{[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2-[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]^2\}\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\phi}_{0j}\Big\|_\infty\geqslant ch(a^2\xi_0)^{-1}\sigma_x\sqrt{6\xi_2}\right)\\
\leqslant &P\left(\max_{1\leqslant i\leqslant n}\big|[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2-[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]^2\big|\geqslant ch\right)\\
&+P\left(\big|\big|\frac{1}{n}\sum_{i=1}^n\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T\big|\big|_\infty\geqslant 3\sigma_x^2\right)+\sum_{j=2}^p P\left(\frac{1}{n}\sum_{i=1}^n(\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\phi}_{0j})^2\geqslant 2\xi_2(a^2\xi_0)^{-2}\right)\\
\leqslant &\exp(-c_1\log p),
\end{aligned}
$$

for some positive constants c, c_1, and all n sufficiently large. □

Lemma B10. *Assume the conditions of Lemma 2 are satisfied, then there exist universal positive constants c_0 and c_1 such that for all n sufficiently large,*

$$
P\left(\max_{2\leqslant j\leqslant p}\Big\|\frac{1}{n}\sum_{i=1}^n[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2(\widehat{\boldsymbol{x}}_{i,-1}\widehat{\boldsymbol{x}}_{i,-1}^T-\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T)\boldsymbol{\phi}_{0j}\Big\|_\infty\geqslant c_0\sqrt{s}h^2\right)\leqslant\exp(-c_1\log p),
$$

where $\boldsymbol{\phi}_{0j}=\tau_{0j}^2\boldsymbol{\theta}_j$, and $\widehat{\boldsymbol{x}}_{i,-1}=\boldsymbol{x}_{i,-1}-\widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})$, $\widetilde{\boldsymbol{x}}_{i,-1}=\boldsymbol{x}_{i,-1}-\mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)$.

Proof. Observe that

$$
\begin{aligned}
&\Big\|\frac{1}{n}\sum_{i=1}^n[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2(\widehat{\boldsymbol{x}}_{i,-1}\widehat{\boldsymbol{x}}_{i,-1}^T-\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T)\boldsymbol{\phi}_{0j}\Big\|_\infty\\
\leqslant&\Big\|\frac{1}{n}\sum_{i=1}^n\left[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\right]^2(\widehat{\boldsymbol{x}}_{i,-1}-\widetilde{\boldsymbol{x}}_{i,-1})(\widehat{\boldsymbol{x}}_{i,-1}-\widetilde{\boldsymbol{x}}_{i,-1})^T\boldsymbol{\phi}_{0j}\Big\|_\infty\\
&+\Big\|\frac{1}{n}\sum_{i=1}^n[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2(\widehat{\boldsymbol{x}}_{i,-1}-\widetilde{\boldsymbol{x}}_{i,-1})\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\phi}_{0j}\Big\|_\infty\\
&+\Big\|\frac{1}{n}\sum_{i=1}^n[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2\widetilde{\boldsymbol{x}}_{i,-1}(\widehat{\boldsymbol{x}}_{i,-1}-\widetilde{\boldsymbol{x}}_{i,-1})^T\boldsymbol{\phi}_{0j}\Big\|_\infty.
\end{aligned}
$$

Inequality (S50) implies that

$$
P\left(\max_{1\leqslant i\leqslant n}\left[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})\right]^2\geqslant b^2+ch\right)\leqslant\exp(-d_1\log p),
$$

for some positive constants c, d_1, and all n sufficiently large. Note that $\widehat{\boldsymbol{x}}_{i,-1} - \widetilde{\boldsymbol{x}}_{i,-1} = \widehat{\mathrm{E}}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}) - \mathrm{E}(\boldsymbol{x}_{i,-1}|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)$. Theorem 1 and Lemma A12 imply that

$$P\left(||\frac{1}{n}\sum_{i=1}^{n}(\widehat{\boldsymbol{x}}_{i,-1} - \widetilde{\boldsymbol{x}}_{i,-1})(\widehat{\boldsymbol{x}}_{i,-1} - \widetilde{\boldsymbol{x}}_{i,-1})^T||_\infty \geqslant c_0 sh^4\right) \leqslant \exp(-c_1\log p),$$
$$P\left(\frac{1}{n}\sum_{i=1}^{n}\left|(\widehat{\boldsymbol{x}}_{i,-1} - \widetilde{\boldsymbol{x}}_{i,-1})^T\boldsymbol{\phi}_{0j}\right|^2 \geqslant c_0 sh^4||\boldsymbol{\phi}_{0j}||_2^2\right) \leqslant \exp(-c_1\log p),$$

for some positive constants c_0, c_1, and all n sufficiently large. Lemma A13 implies that $||\boldsymbol{\phi}_{0j}||_2 = \tau_{0j}^2||\boldsymbol{\theta}_j||_2 \leqslant b^2\xi_1\xi_2^{-1}$. Hence we can conclude that

$$\begin{aligned}
&P\left(\max_{2\leqslant j\leqslant p}\left|\left|\frac{1}{n}\sum_{i=1}^{n}[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2(\widehat{\boldsymbol{x}}_{i,-1} - \widetilde{\boldsymbol{x}}_{i,-1})(\widehat{\boldsymbol{x}}_{i,-1} - \widetilde{\boldsymbol{x}}_{i,-1})^T\boldsymbol{\phi}_{0j}\right|\right|_\infty \geqslant c_0 sh^4\right)\\
\leqslant &P\left(\max_{1\leqslant i\leqslant n}[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2 \geqslant 2b^2\right) + P\left(\left|\left|\frac{1}{n}\sum_{i=1}^{n}(\widehat{\boldsymbol{x}}_{i,-1} - \widetilde{\boldsymbol{x}}_{i,-1})(\widehat{\boldsymbol{x}}_{i,-1} - \widetilde{\boldsymbol{x}}_{i,-1})^T\right|\right|_\infty \geqslant c' sh^4\right)\\
&+\sum_{j=2}^{p}P\left(\frac{1}{n}\sum_{i=1}^{n}\left|(\widehat{\boldsymbol{x}}_{i,-1} - \widetilde{\boldsymbol{x}}_{i,-1})^T\boldsymbol{\phi}_{0j}\right|^2 \geqslant c' b^4\xi_1^2\xi_2^{-2}sh^4\right)\\
\leqslant &\exp(-c_1\log p).
\end{aligned}$$

for some positive constants c_0, c_1, c', and all n sufficiently large. Similarly, we have that $P\left(\max_{2\leqslant j\leqslant p}\left|\left|\frac{1}{n}\sum_{i=1}^{n}[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2(\widehat{\boldsymbol{x}}_{i,-1} - \widetilde{\boldsymbol{x}}_{i,-1})\widetilde{\boldsymbol{x}}_{i,-1}^T\boldsymbol{\phi}_{0j}\right|\right|_\infty \geqslant c_0\sqrt{s}h^2\right) \leqslant \exp(-c_1\log p)$, and $P\left(\max_{2\leqslant j\leqslant p}\left|\left|\frac{1}{n}\sum_{i=1}^{n}[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2\widetilde{\boldsymbol{x}}_{i,-1}(\widehat{\boldsymbol{x}}_{i,-1} - \widetilde{\boldsymbol{x}}_{i,-1})^T\boldsymbol{\phi}_{0j}\right|\right|_\infty \geqslant c_0\sqrt{s}h^2\right) \leqslant \exp(-c_1\log p)$, for some positive constants c_0, c_1, and all n sufficiently large. This concludes the proof of Lemma B10. $\square$

Lemma B11. *Assume the conditions of Lemma 2 are satisfied, then there exist some positive constant c such that for all n sufficiently large,*

$$\max_{2\leqslant j\leqslant p}\widehat{\boldsymbol{\theta}}_j^T\left(\frac{1}{n}\sum_{i=1}\widehat{\boldsymbol{x}}_{i,-1}\widehat{\boldsymbol{x}}_{i,-1}^T\right)\widehat{\boldsymbol{\theta}}_j \leqslant 4\xi_1\xi_2^{-2},$$

with probability at least $1-\exp(-c\log p)$, where ξ_1 and ξ_2 are defined in Assumption (A2)-(a).

Proof. Assumption (A2)-(a) implies that $\inf_{\boldsymbol{v}\in\mathcal{V}_1} \boldsymbol{v}^T\mathrm{E}\big[\mathrm{Cov}(\boldsymbol{x}_{-1}|\boldsymbol{x}^T\boldsymbol{\beta}_0)\big]\boldsymbol{v} \geqslant \xi_0$, where $\mathcal{V}_1 = \{\boldsymbol{v}\in\mathbb{R}^{p-1} : ||\boldsymbol{v}||_2 = 1, ||\boldsymbol{v}||_0 \leqslant 2ks\}$. Define

$$\mathcal{E}_0 = \left\{ \sup_{\boldsymbol{v}\in\mathbb{K}(p-1,2ks)} \left| \boldsymbol{v}^T \Big[\frac{1}{n}\sum_{i=1}^n \widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T - \mathrm{E}(\widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T)\Big]\boldsymbol{v}\right| \geqslant \frac{\xi_0}{54} \right\},$$

where $\mathbb{K}(p-1,2ks) = \{\boldsymbol{v}\in\mathbb{R}^{p-1} : ||\boldsymbol{v}||_0 \leqslant 2ks, ||\boldsymbol{v}||_2 \leqslant 1\}$. By taking $s_0 = \frac{n}{2\log p}$ and $t = \frac{\xi_0}{54}$, Lemma B2 implies that $P\left(\mathcal{E}_0\right) \leqslant 2\exp\left(-cn\min\left\{\frac{\xi_0}{54},1\right\}\right)$, for some positive constant c and all n sufficiently large. Then Lemma 13 in Loh and Wainwright [2012] implies that on the event $\mathcal{E}_0$, for any $\boldsymbol{v}\in\mathbb{R}^{p-1}$, we have that $\boldsymbol{v}^T\big(\frac{1}{n}\sum_{i=1}^n \widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T\big)\boldsymbol{v} \leqslant \frac{3\xi_1}{2}||\boldsymbol{v}||_2^2 + \frac{\xi_0\log p}{n}||\boldsymbol{v}||_1^2$.

Lemma A13 and results in Lemma 2-(3) of the main paper imply that

$$\max_{2\leqslant j\leqslant p} ||\widehat{\boldsymbol{\theta}}_j||_2^2 \leqslant 2\max_{2\leqslant j\leqslant p}(||\boldsymbol{\theta}_j||_2^2 + ||\boldsymbol{\theta}_j - \widehat{\boldsymbol{\theta}}_j||_2^2) \leqslant 2(\xi_2^{-2} + c_0\eta^2\widetilde{s}),$$
$$\max_{2\leqslant j\leqslant p} ||\widehat{\boldsymbol{\theta}}_j||_1^2 \leqslant 2\max_{2\leqslant j\leqslant p}(||\boldsymbol{\theta}_j||_1^2 + ||\boldsymbol{\theta}_j - \widehat{\boldsymbol{\theta}}_j||_1^2) \leqslant 2(\widetilde{s}\xi_2^{-2} + c_0\widetilde{s}^2\eta^2),$$

with probability at least $1-\exp(-c_1\log p)$, for some positive constants c_0, c_1, and all n sufficiently large. Hence on the event $\mathcal{E}_0$,

$$\begin{aligned}
&\max_{2\leqslant j\leqslant p} \widehat{\boldsymbol{\theta}}_j^T \left(\frac{1}{n}\sum_{i=1} \widehat{\boldsymbol{x}}_{i,-1}\widehat{\boldsymbol{x}}_{i,-1}^T\right)\widehat{\boldsymbol{\theta}}_j \\
\leqslant & \max_{2\leqslant j\leqslant p} \widehat{\boldsymbol{\theta}}_j^T \left(\frac{1}{n}\sum_{i=1} \widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T\right)\widehat{\boldsymbol{\theta}}_j + \max_{2\leqslant j\leqslant p} \widehat{\boldsymbol{\theta}}_j^T \left(\frac{1}{n}\sum_{i=1} \widehat{\boldsymbol{x}}_{i,-1}\widehat{\boldsymbol{x}}_{i,-1}^T - \widetilde{\boldsymbol{x}}_{i,-1}\widetilde{\boldsymbol{x}}_{i,-1}^T\right)\widehat{\boldsymbol{\theta}}_j \\
\leqslant & \frac{3\xi_1}{2}\max_{2\leqslant j\leqslant p}||\widehat{\boldsymbol{\theta}}_j||_2^2 + \frac{\xi_0\log p}{n}\max_{2\leqslant j\leqslant p}||\widehat{\boldsymbol{\theta}}_j||_1^2 + 2\max_{2\leqslant j\leqslant p}||\widehat{\boldsymbol{\theta}}_j||_1 \left\|\frac{1}{n}\sum_{i=1}(\widehat{\boldsymbol{x}}_{i,-1} - \widetilde{\boldsymbol{x}}_{i,-1})\widetilde{\boldsymbol{x}}_{i,-1}^T\right\|_\infty \\
& + \max_{2\leqslant j\leqslant p}||\widehat{\boldsymbol{\theta}}_j||_1 \left\|\frac{1}{n}\sum_{i=1}(\widehat{\boldsymbol{x}}_{i,-1} - \widetilde{\boldsymbol{x}}_{i,-1})(\widehat{\boldsymbol{x}}_{i,-1} - \widetilde{\boldsymbol{x}}_{i,-1})^T\right\|_\infty \\
\leqslant & 3\xi_1\left(\xi_2^{-2} + c_0h^2\widetilde{s}\right) + 2\left(\frac{\xi_0\log p}{n} + c_0\sqrt{s}h^2\right)\left(\widetilde{s}\xi_2^{-2} + c_0\widetilde{s}^2h^2\right) \\
\leqslant & 4\xi_1\xi_2^{-2},
\end{aligned}$$

with probability at least $1 - \exp(-c \log p)$, for some positive constants c_0, c, and all n sufficiently large, since $n^{-1}\widetilde{s} \log p = o(h^5 \widetilde{s}) = o(1)$, $\widetilde{s}\sqrt{s} h^2 = o(1)$ and $n^{-1}\widetilde{s}^2 h^2 \log p = o(h^7 \widetilde{s}^2) = o(1)$. In the above, the third inequality applies the results in Lemma B10. Hence it concludes the proof of the lemma. □

S8 Identifiability conditions for the classical low-dimensional single index model

We assume that the underlying low-dimensional true model for the treatment-covariates interaction term $f_0(\boldsymbol{x}^T \boldsymbol{\beta}_0)$ complies with the classical identification assumptions for the single-index model (i.e., our condition (A1)-(c)). To be self-contained, we state below a set of sufficient conditions for identifying $\boldsymbol{\beta}_0$ in the low-dimension model as stated in Theorem 2.1 in Horowitz [2012].

(a) $f_0(\cdot)$ is differentiable and non-constant on the support of $\boldsymbol{x}^T \boldsymbol{\beta}_0$.

(b) The components of $\boldsymbol{x}_{T_0}$ are continuously distributed random variables that have a joint probability density function, where T_0 is the index set corresponding to the nonzero coefficients in $\boldsymbol{\beta}_0$ and $\boldsymbol{x}_{T_0}$ denotes the subvector of $\boldsymbol{x}$ with indices in T_0.

(c) The support of $\boldsymbol{x}_{T_0}$ is not contained in any proper linear space of $\mathbb{R}^s$, with $s = |T_0|$.

(d) $\beta_1 = 1$ and $||\boldsymbol{\beta}_0||_0 \geqslant 2$.

Remark. The literature has slightly different versions of identifiability conditions for the single index model, for example Ichimura [1993]. The above conditions are cited for their transparency. As discussed in Horowitz [2012], a more complex set of conditions are available to allow for some components of $\boldsymbol{x}$ being discrete. In particular, the following two additional conditions are needed: (1) varying the values of the discrete components must not divide the support of $\boldsymbol{x}^T \boldsymbol{\beta}_0$ into disjoint subsets, and (2) $f_0(\cdot)$ must satisfy a non-periodicity condition.

S9 Examples for verifying the regularity conditions

We verify the key conditions on $G(\boldsymbol{x}^T\boldsymbol{\beta}|\boldsymbol{\beta})$, $G^{(1)}(\boldsymbol{x}^T\boldsymbol{\beta}|\boldsymbol{\beta})$, $\mathrm{E}(\boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\beta})$ and $\mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T|\boldsymbol{x}^T\boldsymbol{\beta})$ when $\boldsymbol{x}$ follows a multivariate normal distribution. We focus on conditions that are not much discussed in the current literature on inference for high-dimensional linear regression. For notation simplicity, we assume that $\boldsymbol{x} \sim N(\mathbf{0}, \boldsymbol{I}_p)$. Similar results can be obtained for a multivariate normal distribution with a general covariance $\boldsymbol{\Sigma}$.

Given $\boldsymbol{x} \sim N(\mathbf{0}, \boldsymbol{I}_p)$, then for any $\boldsymbol{w}, \boldsymbol{\beta} \in \mathbb{R}^p$, we have

$$\begin{pmatrix} \boldsymbol{x}^T\boldsymbol{w} \\ \boldsymbol{x}^T\boldsymbol{\beta} \end{pmatrix} \sim N\left(\mathbf{0}, \begin{pmatrix} ||\boldsymbol{w}||_2^2 & \boldsymbol{\beta}^T\boldsymbol{w} \\ \boldsymbol{\beta}^T\boldsymbol{w} & ||\boldsymbol{\beta}||_2^2 \end{pmatrix}\right).$$

For any $\boldsymbol{\beta} \neq \mathbf{0}_p$, the distribution of $\boldsymbol{x}^T\boldsymbol{w}$ conditional on $\boldsymbol{x}^T\boldsymbol{\beta}$ is normal with mean $\frac{\boldsymbol{x}^T\boldsymbol{\beta}}{||\boldsymbol{\beta}||_2^2}\boldsymbol{\beta}^T\boldsymbol{w}$, and variance $||\boldsymbol{w}||_2^2 - \frac{(\boldsymbol{\beta}^T\boldsymbol{w})^2}{||\boldsymbol{\beta}||_2^2}$. We thus have

$$\boldsymbol{x}^T\boldsymbol{\beta}_0|\boldsymbol{x}^T\boldsymbol{\beta} = t \sim N\left(\frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}}{||\boldsymbol{\beta}||_2^2}t, ||\boldsymbol{\beta}_0||_2^2 - \frac{(\boldsymbol{\beta}_0^T\boldsymbol{\beta})^2}{||\boldsymbol{\beta}||_2^2}\right).$$

In the following subsections, we demonstrate the key assumptions in (A2)-(a)(b)(c) and (A5)-(b)(c) hold with high probability in the above setup.

S9.1 Verify Assumption (A2)-(a)

First, we verify the eigenvalue conditions involving $\mathrm{E}[\mathrm{Cov}(\boldsymbol{x}_{-1}|\boldsymbol{x}^T\boldsymbol{\beta}_0)]$ in Assumption (A2)-(a). Note that for any $\boldsymbol{w} \in \mathbb{R}^p$, we have $\mathrm{Var}(\boldsymbol{x}^T\boldsymbol{w}|\boldsymbol{x}^T\boldsymbol{\beta}) = ||\boldsymbol{w}||_2^2 - \frac{(\boldsymbol{\beta}^T\boldsymbol{w})^2}{||\boldsymbol{\beta}||_2^2}$. Recall that $\mathcal{V}_1 = \{\boldsymbol{v} \in \mathbb{R}^{p-1} : ||\boldsymbol{v}||_2 = 1, ||\boldsymbol{v}||_0 \leqslant 2ks\}$, where $k > 1$ is a positive integer. Therefore, we have

$$
\begin{aligned}
\inf_{\boldsymbol{v}\in\mathcal{V}_1} \boldsymbol{v}^T \mathrm{E}\big[\mathrm{Cov}(\boldsymbol{x}_{-1}|\boldsymbol{x}^T\boldsymbol{\beta}_0)\big]\boldsymbol{v} &= \inf_{\boldsymbol{v}\in\mathcal{V}_1} \mathrm{E}\left\{\mathrm{Var}(\boldsymbol{x}_{-1}^T\boldsymbol{v}|\boldsymbol{x}^T\boldsymbol{\beta}_0)\right\} \\
&= \inf_{\boldsymbol{v}\in\mathcal{V}_1} \left[1 - \frac{(\boldsymbol{\beta}_{0,-1}^T\boldsymbol{v})^2}{||\boldsymbol{\beta}_0||_2^2}\right] \\
&\geqslant 1 - \sup_{\boldsymbol{v}\in\mathcal{V}_1} \frac{(\boldsymbol{\beta}_{0,-1}^T\boldsymbol{v})^2}{||\boldsymbol{\beta}_0||_2^2} \\
&\geqslant 1 - \frac{||\boldsymbol{\beta}_{0,-1}||_2^2}{||\boldsymbol{\beta}_0||_2^2} = \frac{1}{||\boldsymbol{\beta}_0||_2^2},
\end{aligned}
$$

since $\boldsymbol{\beta}_0 = (1, \boldsymbol{\beta}_{0,-1}^T)^T$, where the last inequality applies the Cauchy-Schwartz inequality. In the current setup, it is straightforward to show $\lambda_{\max}\left(\mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T)\right) \leqslant 1$. Furthermore,

$$
\begin{aligned}
\lambda_{\max}\left\{\mathrm{E}\big[\mathrm{Cov}(\boldsymbol{x}_{-1}|\boldsymbol{x}^T\boldsymbol{\beta}_0)\big]\right\} &= \sup_{\boldsymbol{v}\in\mathbb{R}^{p-1}:||\boldsymbol{v}||_2=1} \mathrm{E}\left\{\mathrm{Var}(\boldsymbol{x}_{-1}^T\boldsymbol{v}|\boldsymbol{x}^T\boldsymbol{\beta}_0)\right\} \\
&= \sup_{\boldsymbol{v}\in\mathbb{R}^{p-1}:||\boldsymbol{v}||_2=1} \left[1 - \frac{(\boldsymbol{\beta}_{0,-1}^T\boldsymbol{v})^2}{||\boldsymbol{\beta}_0||_2^2}\right] \leqslant 1.
\end{aligned}
$$

The assumption $\lambda_{\min}(\boldsymbol{\Omega}) \geqslant \xi_2$ is similar to the condition imposed on the population Hessian matrix for high-dimensional generalized linear models. To see this is a reasonable assumption, we consider the special case that $\inf_t |f_0'(t)| \geqslant a$ for some positive constant a (e.g., f_0 is a linear function). Then

$$
\begin{aligned}
\lambda_{\min}(\boldsymbol{\Omega}) &= \inf_{\boldsymbol{v}\in\mathbb{R}^{p-1}:||\boldsymbol{v}||_2=1} \mathrm{E}\left\{[G^{(1)}(\boldsymbol{x}^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)]^2(\widetilde{\boldsymbol{x}}_{-1}^T\boldsymbol{v})^2\right\} \\
&\geqslant a^2 \inf_{\boldsymbol{v}\in\mathbb{R}^{p-1}:||\boldsymbol{v}||_2=1} \mathrm{E}\left\{(\widetilde{\boldsymbol{x}}_{-1}^T\boldsymbol{v})^2\right\} \geqslant \frac{a^2}{||\boldsymbol{\beta}_0||_2^2},
\end{aligned}
$$

where the analysis is similar as above, since $\mathrm{E}\left(\widetilde{\boldsymbol{x}}_{-1}\widetilde{\boldsymbol{x}}_{-1}^T\right) = \mathrm{E}\big[\mathrm{Cov}(\boldsymbol{x}_{-1}|\boldsymbol{x}^T\boldsymbol{\beta}_0)\big]$.

Finally, we verify the eigenvalue conditions involving $\lambda_{\max}(\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i^T|\boldsymbol{x}_i^T\boldsymbol{\beta}))$ in Assumption (A2)-(a). For any $\boldsymbol{w} \in \mathbb{R}^p$,

$$\mathrm{E}[(\boldsymbol{x}^T\boldsymbol{w})^2|\boldsymbol{x}^T\boldsymbol{\beta}] = [\mathrm{E}(\boldsymbol{x}^T\boldsymbol{w}|\boldsymbol{x}^T\boldsymbol{\beta})]^2 + \mathrm{Var}(\boldsymbol{x}^T\boldsymbol{w}|\boldsymbol{x}^T\boldsymbol{\beta}) = \frac{(\boldsymbol{x}^T\boldsymbol{\beta})^2(\boldsymbol{\beta}^T\boldsymbol{w})^2}{||\boldsymbol{\beta}||_2^4} + ||\boldsymbol{w}||_2^2 - \frac{(\boldsymbol{\beta}^T\boldsymbol{w})^2}{||\boldsymbol{\beta}||_2^2}.$$

Therefore, we have

$$\begin{aligned}
&\sup_{\boldsymbol{\beta}\in\mathbb{B}} n^{-1}\sum_{i=1}^n \left[\lambda_{\max}\big(\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i^T|\boldsymbol{x}_i^T\boldsymbol{\beta})\big)\right]^2 \\
=&\sup_{\boldsymbol{\beta}\in\mathbb{B}} n^{-1}\sum_{i=1}^n \sup_{\boldsymbol{v}\in\mathbb{R}^p:||\boldsymbol{v}||_2=1} \left\{\mathrm{E}[(\boldsymbol{x}_i^T\boldsymbol{v})^2|\boldsymbol{x}_i^T\boldsymbol{\beta}]\right\}^2 \\
=&\sup_{\boldsymbol{\beta}\in\mathbb{B}} n^{-1}\sum_{i=1}^n \sup_{\boldsymbol{v}\in\mathbb{R}^p:||\boldsymbol{v}||_2=1} \left[\frac{(\boldsymbol{x}_i^T\boldsymbol{\beta})^2(\boldsymbol{\beta}^T\boldsymbol{v})^2}{||\boldsymbol{\beta}||_2^4} + 1 - \frac{(\boldsymbol{\beta}^T\boldsymbol{v})^2}{||\boldsymbol{\beta}||_2^2}\right]^2 \\
\leqslant&\sup_{\boldsymbol{\beta}\in\mathbb{B}} n^{-1}\sum_{i=1}^n \sup_{\boldsymbol{v}\in\mathbb{R}^p:||\boldsymbol{v}||_2=1} \left[\frac{2(\boldsymbol{x}_i^T\boldsymbol{\beta})^4(\boldsymbol{\beta}^T\boldsymbol{v})^4}{||\boldsymbol{\beta}||_2^8} + 2\right] \\
\leqslant&\sup_{\boldsymbol{\beta}\in\mathbb{B}} n^{-1}\sum_{i=1}^n \left[\frac{2(\boldsymbol{x}_i^T\boldsymbol{\beta})^4}{||\boldsymbol{\beta}||_2^4} + 2\right] \\
\leqslant&\sup_{\boldsymbol{\beta}\in\mathbb{B}} \left\{\frac{3\mathrm{E}\left[(\boldsymbol{x}_i^T\boldsymbol{\beta})^4\right]}{||\boldsymbol{\beta}||_2^4} + 2\right\} = 11,
\end{aligned}$$

with probability at least $1 - \exp(-c_1\sqrt{n})$, for some positive constants c_0, c_1, and all n sufficiently large. In the above, the second last inequality applies the Cauchy-Schwartz inequality,and the last inequality applies Lemma B3. Similarly, we have

$$\begin{aligned}
&\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1} \lambda_{\max}\big(\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i^T|\boldsymbol{x}_i^T\boldsymbol{\beta})\big) \\
=&\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\sup_{\boldsymbol{v}\in\mathbb{R}^p:||\boldsymbol{v}||_2=1} \mathrm{E}[(\boldsymbol{x}_i^T\boldsymbol{v})^2|\boldsymbol{x}_i^T\boldsymbol{\beta}] \\
=&\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\sup_{\boldsymbol{v}\in\mathbb{R}^p:||\boldsymbol{v}||_2=1} \left[\frac{(\boldsymbol{x}_i^T\boldsymbol{\beta})^2(\boldsymbol{\beta}^T\boldsymbol{v})^2}{||\boldsymbol{\beta}||_2^4} + 1 - \frac{(\boldsymbol{\beta}^T\boldsymbol{v})^2}{||\boldsymbol{\beta}||_2^2}\right] \\
\leqslant&\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1} \frac{(\boldsymbol{x}_i^T\boldsymbol{\beta})^2}{||\boldsymbol{\beta}||_2^2} + 1,
\end{aligned}$$

where the last inequality applies the Cauchy-Schwartz inequality. Since $\boldsymbol{x}_i^T\boldsymbol{\beta} \sim N(0, ||\boldsymbol{\beta}||_2^2)$, by the tail property of the normal distribution, we have $P\Big(\max_{1\leqslant i\leqslant n}|\boldsymbol{x}_i^T\boldsymbol{\beta}|\geqslant c_0||\boldsymbol{\beta}||_2\sqrt{\log(p\vee n)}$, $\forall\boldsymbol{\beta}\in\mathbb{B}_1\Big)\leqslant\exp[-c_1\log(p\vee n)]$, some positive constants c_0, c_1, and all n sufficiently large. Thus we have

$$
\begin{aligned}
&P\left(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\lambda_{\max}\big(\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i^T|\boldsymbol{x}_i^T\boldsymbol{\beta})\big)\geqslant M\log(p\vee n)\right)\\
\leqslant &P\left(\max_{1\leqslant i\leqslant n}(\boldsymbol{x}_i^T\boldsymbol{\beta})^2\geqslant[M\log(p\vee n)-1]||\boldsymbol{\beta}||_2^2,\ \forall\ \boldsymbol{\beta}\in\mathbb{B}_1\right)\\
\leqslant &\exp[-c_1\log(p\vee n)],
\end{aligned}
$$

for some positive constants M, c_1, and all n sufficiently large.

S9.2 Verify Assumption (A2)-(b)

Next we verify the key conditions in (A2)-(b). Observe that $\mathrm{E}(\boldsymbol{x}_{-1}^T\boldsymbol{\eta}|\boldsymbol{x}^T\boldsymbol{\beta}=t)=\frac{t}{||\boldsymbol{\beta}||_2^2}\boldsymbol{\beta}_{-1}^T\boldsymbol{\eta}$. Hence we have $\mathrm{E}^{(1)}(\boldsymbol{x}_{-1}^T\boldsymbol{\eta}|\boldsymbol{x}^T\boldsymbol{\beta}=t)=\frac{\boldsymbol{\beta}_{-1}^T\boldsymbol{\eta}}{||\boldsymbol{\beta}||_2^2}$, and $\mathrm{E}^{(2)}(\boldsymbol{x}_{-1}^T\boldsymbol{\eta}|\boldsymbol{x}^T\boldsymbol{\beta}=t)=0$, satisfying the following constraints:

$$
\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}}|\mathrm{E}^{(1)}(\boldsymbol{x}_{i,-1}^T\boldsymbol{\eta}|\boldsymbol{x}_i^T\boldsymbol{\beta})|=\sup_{\boldsymbol{\beta}\in\mathbb{B}}\frac{|\boldsymbol{\beta}_{-1}^T\boldsymbol{\eta}|}{||\boldsymbol{\beta}||_2^2}\leqslant\sup_{\boldsymbol{\beta}\in\mathbb{B}}\frac{||\boldsymbol{\eta}||_2}{||\boldsymbol{\beta}||_2}\leqslant||\boldsymbol{\eta}||_2,
$$

$$
\sup_{|t|\leqslant 2||\boldsymbol{\beta}_0||_2\sigma_x\sqrt{\log(p\vee n)}}\sup_{\boldsymbol{\beta}\in\mathbb{B}}|\mathrm{E}^{(2)}(\boldsymbol{x}_{-1}^T\boldsymbol{\eta}|\boldsymbol{x}^T\boldsymbol{\beta}=t)|\leqslant||\boldsymbol{\eta}||_2,
$$

since $||\boldsymbol{\beta}||_2\geqslant 1$ for any $\boldsymbol{\beta}\in\mathbb{B}$. Recall that $\mathbb{B}_1=\{\boldsymbol{\beta}\in\mathbb{B}:||\boldsymbol{\beta}-\boldsymbol{\beta}_0||_2\leqslant c_0\sqrt{s}h^2,||\boldsymbol{\beta}||_0\leqslant ks\}$, for some constants $k>1$ and $c_0>0$. Since $\mathrm{E}[(\boldsymbol{x}_{-1}^T\boldsymbol{\eta})^2|\boldsymbol{x}^T\boldsymbol{\beta}=t]=\frac{t^2}{||\boldsymbol{\beta}||_2^2}\boldsymbol{\beta}_{-1}^T\boldsymbol{\eta}+||\boldsymbol{\eta}||_2^2-\frac{(\boldsymbol{\beta}_{-1}^T\boldsymbol{\eta})^2}{||\boldsymbol{\beta}||_2^2}$, the Cauchy-Schwartz inequality implies that

$$
\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left\{\left|\mathrm{E}^{(1)}\left[(\boldsymbol{x}_{i,-1}^T\boldsymbol{\eta})^2|\boldsymbol{x}_i^T\boldsymbol{\beta}\right]\right|\right\}=\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\frac{2|\boldsymbol{x}_i^T\boldsymbol{\beta}|(\boldsymbol{\beta}_{-1}^T\boldsymbol{\eta})^2}{||\boldsymbol{\beta}||_2^4}\leqslant||\boldsymbol{\eta}||_2^2\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\frac{2|\boldsymbol{x}_i^T\boldsymbol{\beta}|}{||\boldsymbol{\beta}||_2^2}.
$$

We have $P\left(\max_{1\leqslant i\leqslant n}|\boldsymbol{x}_i^T\boldsymbol{\beta}|\geqslant c_0||\boldsymbol{\beta}||_2\sqrt{\log(p\vee n)},\ \forall\ \boldsymbol{\beta}\in\mathbb{B}_1\right)\leqslant\exp[-c\log(p\vee n)]$, for some positive constants c_0, c_1, and all n sufficiently large. Note that for any $\boldsymbol{\beta}\in\mathbb{B}_1$, we have $\beta_1 = 1$. We thus have

$$1 \leqslant ||\boldsymbol{\beta}||_2 \leqslant ||\boldsymbol{\beta}_0||_2 + ||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_2 \leqslant ||\boldsymbol{\beta}_0||_2 + c_0\sqrt{s}h^2 \leqslant c_0,$$

for some constant $c_0 > 0$. Then we have

$$\begin{aligned}
&P\left(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\left\{\left|\mathrm{E}^{(1)}\left[(\boldsymbol{x}_{i,-1}^T\boldsymbol{\eta})^2|\boldsymbol{x}_i^T\boldsymbol{\beta}\right]\right|\right\} \geqslant M||\boldsymbol{\eta}||_2^2\sqrt{\log(p\vee n)}\right)\\
\leqslant &P\left(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}\frac{2|\boldsymbol{x}_i^T\boldsymbol{\beta}|}{||\boldsymbol{\beta}||_2^2} \geqslant M\sqrt{\log(p\vee n)}\right)\\
\leqslant &P\left(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}|\boldsymbol{x}_i^T\boldsymbol{\beta}| \geqslant c_0\sqrt{\log(p\vee n)}\right)\\
\leqslant &\exp[-c_1\log(p\vee n)],
\end{aligned}$$

for some positive constants M, c_0, c_1, and all n sufficiently large.

S9.3 Verify the Lipschitz condition of $\mathbf{E}(\boldsymbol{x}|\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{\beta})$ in Assumption (A2)-(c)

For any $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{B}$, we observe that

$$\begin{aligned}
\mathrm{E}(\boldsymbol{x}^T\boldsymbol{v}|\boldsymbol{x}^T\boldsymbol{\beta}_1)-\mathrm{E}(\boldsymbol{x}^T\boldsymbol{v}|\boldsymbol{x}^T\boldsymbol{\beta}_2) &= \frac{\boldsymbol{x}^T\boldsymbol{\beta}_1}{||\boldsymbol{\beta}_1||_2^2}\boldsymbol{\beta}_1^T\boldsymbol{v} - \frac{\boldsymbol{x}^T\boldsymbol{\beta}_2}{||\boldsymbol{\beta}_2||_2^2}\boldsymbol{\beta}_2^T\boldsymbol{v}\\
&=\frac{\boldsymbol{\beta}_1^T\boldsymbol{v}}{||\boldsymbol{\beta}_1||_2^2}(\boldsymbol{x}^T\boldsymbol{\beta}_1 - \boldsymbol{x}^T\boldsymbol{\beta}_2) + \boldsymbol{x}^T\boldsymbol{\beta}_2\left(\frac{\boldsymbol{\beta}_1^T\boldsymbol{v}}{||\boldsymbol{\beta}_1||_2^2} - \frac{\boldsymbol{\beta}_2^T\boldsymbol{v}}{||\boldsymbol{\beta}_2||_2^2}\right)\\
&=\frac{\boldsymbol{\beta}_1^T\boldsymbol{v}}{||\boldsymbol{\beta}_1||_2^2}(\boldsymbol{x}^T\boldsymbol{\beta}_1 - \boldsymbol{x}^T\boldsymbol{\beta}_2) + (\boldsymbol{x}^T\boldsymbol{\beta}_2)\frac{(\boldsymbol{\beta}_1-\boldsymbol{\beta}_2)^T\boldsymbol{v}}{||\boldsymbol{\beta}_1||_2^2} + (\boldsymbol{x}^T\boldsymbol{\beta}_2)(\boldsymbol{\beta}_2^T\boldsymbol{v})\left(\frac{1}{||\boldsymbol{\beta}_1||_2^2} - \frac{1}{||\boldsymbol{\beta}_2||_2^2}\right)\\
&\triangleq A_1 + A_2 + A_3,
\end{aligned}$$

where the definition of $A_k,\ k = 1, \cdots, 3$, is clear from the context. Note that for the identifiability condition assumes $\beta_1 = 1$. We have

$$1 \leqslant ||\boldsymbol{\beta}||_2 \leqslant ||\boldsymbol{\beta}_0||_2 + ||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_2 \leqslant ||\boldsymbol{\beta}_0||_2 + r \leqslant c_0, \tag{S51}$$

for some constant $c_0 > 0$. By the Cauchy-Schwartz Inequality and (S51), we obtain that $\sup_{\boldsymbol{v}\in\mathbb{K}(2ks+\widetilde{s})} |A_1| \leqslant \sup_{\boldsymbol{v}\in\mathbb{K}(2ks+\widetilde{s})} |\boldsymbol{x}^T\boldsymbol{\beta}_1 - \boldsymbol{x}^T\boldsymbol{\beta}_2| * \frac{||\boldsymbol{\beta}_1||_2||\boldsymbol{v}||_2}{||\boldsymbol{\beta}_1||_2^2} \leqslant |\boldsymbol{x}^T\boldsymbol{\beta}_1 - \boldsymbol{x}^T\boldsymbol{\beta}_2|$, and $\sup_{\boldsymbol{v}\in\mathbb{K}(2ks+\widetilde{s})} |A_2| \leqslant |\boldsymbol{x}^T\boldsymbol{\beta}_2| * ||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2||_2$. To bound $\sup_{\boldsymbol{v}\in\mathbb{K}(2ks+\widetilde{s})} |A_3|$, observe that

$$\begin{aligned}
\sup_{\boldsymbol{v}\in\mathbb{K}(2ks+\widetilde{s})} |A_3| &= \sup_{\boldsymbol{v}\in\mathbb{K}(2ks+\widetilde{s})} |\boldsymbol{x}^T\boldsymbol{\beta}_2| * |\boldsymbol{\beta}_2^T\boldsymbol{v}| * \left|\frac{||\boldsymbol{\beta}_2||_2^2 - ||\boldsymbol{\beta}_1||_2^2}{||\boldsymbol{\beta}_1||_2^2||\boldsymbol{\beta}_2||_2^2}\right| \\
&= \sup_{\boldsymbol{v}\in\mathbb{K}(2ks+\widetilde{s})} |\boldsymbol{x}^T\boldsymbol{\beta}_2| * |\boldsymbol{\beta}_2^T\boldsymbol{v}| * \frac{\big|(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2)^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)\big|}{||\boldsymbol{\beta}_1||_2^2||\boldsymbol{\beta}_2||_2^2} \\
&\leqslant \sup_{\boldsymbol{v}\in\mathbb{K}(2ks+\widetilde{s})} |\boldsymbol{x}^T\boldsymbol{\beta}_2| * ||\boldsymbol{\beta}_2||_2 * ||\boldsymbol{v}||_2 * \frac{||\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2||_2||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2||_2}{||\boldsymbol{\beta}_1||_2^2||\boldsymbol{\beta}_2||_2^2} \\
&\leqslant |\boldsymbol{x}^T\boldsymbol{\beta}_2| * ||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2||_2 * \frac{||\boldsymbol{\beta}_1||_2 + ||\boldsymbol{\beta}_2||_2}{||\boldsymbol{\beta}_1||_2^2||\boldsymbol{\beta}_2||_2} \\
&\leqslant 2|\boldsymbol{x}^T\boldsymbol{\beta}_2| * ||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2||_2,
\end{aligned}$$

where the last inequality applies (S51). Combining all these results, we show that

$$\sup_{\boldsymbol{v}\in\mathbb{K}(2ks+\widetilde{s})} \Big|\mathrm{E}(\boldsymbol{x}^T\boldsymbol{v}|\boldsymbol{x}^T\boldsymbol{\beta}_1) - \mathrm{E}(\boldsymbol{x}^T\boldsymbol{v}|\boldsymbol{x}^T\boldsymbol{\beta}_2)\Big| \leqslant 3\big(|\boldsymbol{x}^T\boldsymbol{\beta}_1 - \boldsymbol{x}^T\boldsymbol{\beta}_2| + |\boldsymbol{x}^T\boldsymbol{\beta}_2| * ||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2||_2\big).$$

Similarly, we can also show that

$$\sup_{\boldsymbol{v}\in\mathbb{K}(2ks+\widetilde{s})} \Big|\mathrm{E}(\boldsymbol{x}^T\boldsymbol{v}|\boldsymbol{x}^T\boldsymbol{\beta}_1) - \mathrm{E}(\boldsymbol{x}^T\boldsymbol{v}|\boldsymbol{x}^T\boldsymbol{\beta}_2)\Big| \leqslant 3\big(|\boldsymbol{x}^T\boldsymbol{\beta}_1 - \boldsymbol{x}^T\boldsymbol{\beta}_2| + |\boldsymbol{x}^T\boldsymbol{\beta}_1| * ||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2||_2\big).$$

Hence, it implies that

$$\sup_{\boldsymbol{v}\in\mathbb{K}(2ks+\widetilde{s})} \Big|\mathrm{E}(\boldsymbol{x}^T\boldsymbol{v}|\boldsymbol{x}^T\boldsymbol{\beta}_1) - \mathrm{E}(\boldsymbol{x}^T\boldsymbol{v}|\boldsymbol{x}^T\boldsymbol{\beta}_2)\Big| \leqslant 3\big[|\boldsymbol{x}^T\boldsymbol{\beta}_1 - \boldsymbol{x}^T\boldsymbol{\beta}_2| + \min(|\boldsymbol{x}^T\boldsymbol{\beta}_1|, |\boldsymbol{x}^T\boldsymbol{\beta}_2|) * ||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2||_2\big].$$

S9.4 Verify the assumptions on $\mathrm{G}(\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{\beta}|\boldsymbol{\beta})$ in (A5)-(c)

Let $\sigma_{\boldsymbol{\beta}}^2 = ||\boldsymbol{\beta}_0||_2^2 - \frac{(\boldsymbol{\beta}_0^T\boldsymbol{\beta})^2}{||\boldsymbol{\beta}||_2^2}$, and $\phi(\cdot)$ be the p.d.f of $N(0,1)$. We observe that

$$G(t|\boldsymbol{\beta}) = \mathrm{E}[f_0(\boldsymbol{x}^T\boldsymbol{\beta}_0)|\boldsymbol{x}^T\boldsymbol{\beta} = t] = \int f_0(z)\sigma_{\boldsymbol{\beta}}^{-1}\phi\left(\frac{z - \frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}}{||\boldsymbol{\beta}||_2^2}t}{\sigma_{\boldsymbol{\beta}}}\right)dz.$$

Let $w = \sigma_{\boldsymbol{\beta}}^{-1}\left(z - \frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}}{||\boldsymbol{\beta}||_2^2}t\right)$, by a transformation of variable, we have

$$G(t|\boldsymbol{\beta}) = \int f_0\left(\sigma_{\boldsymbol{\beta}}w + \frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}}{||\boldsymbol{\beta}||_2^2}t\right)\phi(w)dw. \tag{S52}$$

Let $\sigma_{\boldsymbol{\beta}_1}^2 = ||\boldsymbol{\beta}_0||_2^2 - \frac{(\boldsymbol{\beta}_0^T\boldsymbol{\beta}_1)^2}{||\boldsymbol{\beta}_1||_2^2}$, and $\sigma_{\boldsymbol{\beta}_2}^2 = ||\boldsymbol{\beta}_0||_2^2 - \frac{(\boldsymbol{\beta}_0^T\boldsymbol{\beta}_2)^2}{||\boldsymbol{\beta}_2||_2^2}$. Then we have

$$\begin{aligned} G(t|\boldsymbol{\beta}_1) - G(t|\boldsymbol{\beta}_2) &= \int\left[f_0\Big(\sigma_{\boldsymbol{\beta}_1}w + \frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}_1}{||\boldsymbol{\beta}_1||_2^2}t\Big) - f_0\Big(\sigma_{\boldsymbol{\beta}_2}w + \frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}_2}{||\boldsymbol{\beta}_2||_2^2}t\Big)\right]\phi(w)dw \\ &= \int f_0'(\widetilde{w})\left[(\sigma_{\boldsymbol{\beta}_1} - \sigma_{\boldsymbol{\beta}_2})w + \frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}_1}{||\boldsymbol{\beta}_1||_2^2}t - \frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}_2}{||\boldsymbol{\beta}_2||_2^2}t\right]\phi(w)dw, \end{aligned} \tag{S53}$$

where $\widetilde{w}$ is between $\sigma_{\boldsymbol{\beta}_1}w + \frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}_1}{||\boldsymbol{\beta}_1||_2^2}t$ and $\sigma_{\boldsymbol{\beta}_2}w + \frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}_2}{||\boldsymbol{\beta}_2||_2^2}t$. Assumption (A1)-(b) indicates that f_0 is differentiable and $\max_{1\leqslant i\leqslant n}|f_0'(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)| \leqslant b$. Then we can obtain

$$|G(\boldsymbol{x}^T\boldsymbol{\beta}_1|\boldsymbol{\beta}_1) - G(\boldsymbol{x}^T\boldsymbol{\beta}_2|\boldsymbol{\beta}_2)| \leqslant b|\sigma_{\boldsymbol{\beta}_1} - \sigma_{\boldsymbol{\beta}_2}| * \mathrm{E}|w| + b|t| * \left|\frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}_1}{||\boldsymbol{\beta}_1||_2^2} - \frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}_2}{||\boldsymbol{\beta}_2||_2^2}\right|.$$

As $w \sim N(0,1)$, we have $\mathrm{E}|w| = \sqrt{2/\pi}$. According to analysis in Section S9.3, we have that $\left|\frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}_1}{||\boldsymbol{\beta}_1||_2^2} - \frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}_2}{||\boldsymbol{\beta}_2||_2^2}\right| \leqslant c_1||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2||_2$, for some positive constant c_1. Without loss of generality, we assume $\sigma_{\boldsymbol{\beta}_1} \geqslant \sigma_{\boldsymbol{\beta}_2} > 0$. then $|\sigma_{\boldsymbol{\beta}_1}^2 - \sigma_{\boldsymbol{\beta}_2}^2| = (\sigma_{\boldsymbol{\beta}_1} - \sigma_{\boldsymbol{\beta}_2})^2 + 2\sigma_{\boldsymbol{\beta}_2}(\sigma_{\boldsymbol{\beta}_1} - \sigma_{\boldsymbol{\beta}_2}) \geqslant (\sigma_{\boldsymbol{\beta}_1} - \sigma_{\boldsymbol{\beta}_2})^2$. We thus have

$$
\begin{aligned}
|\sigma_{\boldsymbol{\beta}_1} - \sigma_{\boldsymbol{\beta}_2}| \leqslant \sqrt{|\sigma^2_{\boldsymbol{\beta}_1} - \sigma^2_{\boldsymbol{\beta}_2}|} &= \sqrt{\frac{(\boldsymbol{\beta}_0^T\boldsymbol{\beta}_1)^2}{||\boldsymbol{\beta}_1||_2^2} - \frac{(\boldsymbol{\beta}_0^T\boldsymbol{\beta}_2)^2}{||\boldsymbol{\beta}_2||_2^2}} \\
&\leqslant \sqrt{\frac{(\boldsymbol{\beta}_0^T\boldsymbol{\beta}_1)^2 - (\boldsymbol{\beta}_0^T\boldsymbol{\beta}_2)^2}{||\boldsymbol{\beta}_1||_2^2} + (\boldsymbol{\beta}_0^T\boldsymbol{\beta}_2)^2\left(\frac{1}{||\boldsymbol{\beta}_1||_2^2} - \frac{1}{||\boldsymbol{\beta}_2||_2^2}\right)} \\
&\leqslant \sqrt{\frac{\boldsymbol{\beta}_0^T(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) * \boldsymbol{\beta}_0^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)}{||\boldsymbol{\beta}_1||_2^2}} + \frac{|\boldsymbol{\beta}_0^T\boldsymbol{\beta}_2|}{||\boldsymbol{\beta}_1||_2||\boldsymbol{\beta}_2||_2}\sqrt{(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2)^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)} \\
&\leqslant c_1||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2||_2^{1/2},
\end{aligned}
$$

for some positive constant c_1, where the last inequality applies (S51) and Assumption (A1)-(b). Note that $||\boldsymbol{\beta}_0||_2$ is bounded by Assumption (A1)-(b). Combining all these results, we conclude that for some constant $C > 0$,

$$
|G(t|\boldsymbol{\beta}_1) - G(t|\boldsymbol{\beta}_2)| \leqslant C\Big(|t| * ||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2||_2 + \sqrt{||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2||_2}\Big).
$$

Hence we have

$$
\begin{aligned}
\sup_{|t|\leqslant c_0\sqrt{s\log(p\vee n)}} \big[G(t|\boldsymbol{\beta}_1) - G(t|\boldsymbol{\beta}_2)\big]^2 \leqslant & \sup_{|t|\leqslant c_0\sqrt{s\log(p\vee n)}} 2C^2\Big(t^2 * ||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2||_2^2 + ||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2||_2\Big) \\
\leqslant & 2C^2\Big(2rs\log(p\vee n)||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2||_2 + ||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2||_2\Big) \\
\leqslant & c_1 s\log(p\vee n)||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2||_2,
\end{aligned}
$$

for some positive constants C, c_1, and all n sufficiently large. We thus have validated the assumption on $G(\cdot|\boldsymbol{\beta})$ in (A5)-(c).

S9.5 Verify the assumptions on $\mathbf{G}^{(1)}(\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{\beta}|\boldsymbol{\beta})$ in (A5)-(b) and (A5)-(c)

To validate the assumption on $G^{(1)}(\boldsymbol{x}^T\boldsymbol{\beta}|\boldsymbol{\beta})$ in (A5)-(b), we first note that

$$G^{(1)}(t|\boldsymbol{\beta}) = \frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}}{||\boldsymbol{\beta}||_2^2}\int f_0'\Big(\sigma_{\boldsymbol{\beta}}w + \frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}}{||\boldsymbol{\beta}||_2^2}t\Big)\phi(w)dw,$$

by (S52), where $\sigma_{\boldsymbol{\beta}}^2 = ||\boldsymbol{\beta}_0||_2^2 - \frac{(\boldsymbol{\beta}_0^T\boldsymbol{\beta})^2}{||\boldsymbol{\beta}||_2^2}$.

By Taylor expansion, for some $\widetilde{w}$ between $\sigma_{\boldsymbol{\beta}}w + \frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}}{||\boldsymbol{\beta}||_2^2}\boldsymbol{x}^T\boldsymbol{\beta}$ and $\boldsymbol{x}^T\boldsymbol{\beta}_0$,

$$\begin{aligned}
&G^{(1)}(\boldsymbol{x}^T\boldsymbol{\beta}|\boldsymbol{\beta}) - G^{(1)}(\boldsymbol{x}^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\\
=&\int\left[\frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}}{||\boldsymbol{\beta}||_2^2}f_0'\Big(\sigma_{\boldsymbol{\beta}}w + \frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}}{||\boldsymbol{\beta}||_2^2}\boldsymbol{x}^T\boldsymbol{\beta}\Big) - f_0'(\boldsymbol{x}^T\boldsymbol{\beta}_0)\right]\phi(w)dw\\
=&\frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}}{||\boldsymbol{\beta}||_2^2}\int\left[f_0'\Big(\sigma_{\boldsymbol{\beta}}w + \frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}}{||\boldsymbol{\beta}||_2^2}\boldsymbol{x}^T\boldsymbol{\beta}\Big) - f_0'(\boldsymbol{x}^T\boldsymbol{\beta}_0)\right]\phi(w)dw + f_0'(\boldsymbol{x}^T\boldsymbol{\beta}_0)\left(\frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}}{||\boldsymbol{\beta}||_2^2} - 1\right)\\
\triangleq& D_1 + D_2,
\end{aligned}$$

where the definitions of D_1 and D_2 are clear from the context. By the assumptions for $f_0(\cdot)$ in Assumption (A1)-(b), we have $|D_2| \leqslant b * \frac{|\boldsymbol{\beta}^T(\boldsymbol{\beta}_0-\boldsymbol{\beta})|}{||\boldsymbol{\beta}||_2^2} \leqslant b||\boldsymbol{\beta}-\boldsymbol{\beta}_0||_2$, by (S51).

To bound $|D_1|$, by Taylor expansion and Assumption (A1)-(b), there exist some positive constants c, c_1, c_2, such that

$$\begin{aligned}
|D_1| \leqslant&\left|c\int f_0''(\boldsymbol{x}^T\boldsymbol{\beta}_0)\left(\sigma_{\boldsymbol{\beta}}w + \frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}}{||\boldsymbol{\beta}||_2^2}\boldsymbol{x}^T\boldsymbol{\beta} - \boldsymbol{x}^T\boldsymbol{\beta}_0\right)\phi(w)dw\right|\\
&+ c_1\int\left(\sigma_{\boldsymbol{\beta}}w + \frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}}{||\boldsymbol{\beta}||_2^2}\boldsymbol{x}^T\boldsymbol{\beta} - \boldsymbol{x}^T\boldsymbol{\beta}_0\right)^2\phi(w)dw\\
\leqslant& c_2\left|\frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}}{||\boldsymbol{\beta}||_2^2}\boldsymbol{x}^T\boldsymbol{\beta} - \boldsymbol{x}^T\boldsymbol{\beta}_0\right| + c_2\sigma_{\boldsymbol{\beta}}^2 + c_2\left(\frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}}{||\boldsymbol{\beta}||_2^2}\boldsymbol{x}^T\boldsymbol{\beta} - \boldsymbol{x}^T\boldsymbol{\beta}_0\right)^2.
\end{aligned}$$

Similarly as in Section S9.4, we can obtain that

$$\left|\frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}}{||\boldsymbol{\beta}||_2^2}\boldsymbol{x}^T\boldsymbol{\beta}-\boldsymbol{x}^T\boldsymbol{\beta}_0\right|\leqslant c||\boldsymbol{\beta}_0||_2 * \left(|\boldsymbol{x}^T\boldsymbol{\beta}-\boldsymbol{x}^T\boldsymbol{\beta}_0|+|\boldsymbol{x}^T\boldsymbol{\beta}_0| * ||\boldsymbol{\beta}-\boldsymbol{\beta}_0||_2\right),$$
$$\text{and} \qquad \sigma_{\boldsymbol{\beta}}^2=||\boldsymbol{\beta}_0||_2^2-\frac{(\boldsymbol{\beta}_0^T\boldsymbol{\beta})^2}{||\boldsymbol{\beta}||_2^2}\leqslant ||\boldsymbol{\beta}_0||_2^2 * ||\boldsymbol{\beta}-\boldsymbol{\beta}_0||_2.$$

Then we have for some positive constant C,

$$\begin{aligned}\left|G^{(1)}(\boldsymbol{x}^T\boldsymbol{\beta}|\boldsymbol{\beta})-G^{(1)}(\boldsymbol{x}^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\right|\leqslant& C\Big[|\boldsymbol{x}^T\boldsymbol{\beta}-\boldsymbol{x}^T\boldsymbol{\beta}_0|+|\boldsymbol{x}^T\boldsymbol{\beta}-\boldsymbol{x}^T\boldsymbol{\beta}_0|^2\\&+(1+|\boldsymbol{x}^T\boldsymbol{\beta}_0|) * ||\boldsymbol{\beta}-\boldsymbol{\beta}_0||_2+|\boldsymbol{x}^T\boldsymbol{\beta}_0|^2 * ||\boldsymbol{\beta}-\boldsymbol{\beta}_0||_2^2\Big].\end{aligned}$$

Note that $||\boldsymbol{\beta}-\boldsymbol{\beta}_0||_2^2\leqslant 2$. We have

$$\begin{aligned}&\sup_{\boldsymbol{\beta}\in\mathbb{B}} n^{-1}\sum_{i=1}^n ||\boldsymbol{\beta}-\boldsymbol{\beta}_0||_2^{-2}\big[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0|\boldsymbol{\beta}_0)\big]^2\\ \leqslant& c_0\sup_{\boldsymbol{\beta}\in\mathbb{B}} n^{-1}\sum_{i=1}^n\Big[\frac{|\boldsymbol{x}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)|^2}{||\boldsymbol{\beta}-\boldsymbol{\beta}_0||_2^2}+\frac{|\boldsymbol{x}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)|^4}{||\boldsymbol{\beta}-\boldsymbol{\beta}_0||_2^2}+1+|\boldsymbol{x}^T\boldsymbol{\beta}_0|^2+|\boldsymbol{x}^T\boldsymbol{\beta}_0|^4 * ||\boldsymbol{\beta}-\boldsymbol{\beta}_0||_2^2\Big]\\ \leqslant& c_1,\end{aligned}$$

with probability at least $1-\exp[-c_2 s\log(p\vee n)]$, for some positive constants c_0, c_1, c_2 and all n sufficiently large. In the above, the first part of the inequality applies Lemma B2, the second part of the inequality applies Lemma B3, and the remaining part applies Lemma B1 and Lemma B3. Then we validate Assumption (A5)-(b).

Next we validate assumptions for $G^{(1)}(\boldsymbol{x}^T\boldsymbol{\beta}|\boldsymbol{\beta})$ in (A5)-(c). By Assumption (A1)-(b), for some positive constant c_1,

$$
\begin{aligned}
&\left|G^{(1)}(t|\boldsymbol{\beta}_1) - G^{(1)}(t|\boldsymbol{\beta}_2)\right| \\
&\leqslant \left|\frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}_1}{||\boldsymbol{\beta}_1||_2^2} - \frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}_2}{||\boldsymbol{\beta}_2||_2^2}\right| \int f_0'\left(\sigma_{\boldsymbol{\beta}_1}w + \frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}_1}{||\boldsymbol{\beta}_1||_2^2}t\right)\phi(w)dw \\
&+ \frac{|\boldsymbol{\beta}_0^T\boldsymbol{\beta}_2|}{||\boldsymbol{\beta}_2||_2^2}\int\left|f_0'\left(\sigma_{\boldsymbol{\beta}_1}w + \frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}_1}{||\boldsymbol{\beta}_1||_2^2}t\right) - f_0'\left(\sigma_{\boldsymbol{\beta}_2}w + \frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}_2}{||\boldsymbol{\beta}_2||_2^2}t\right)\right|\phi(w)dw \\
\leqslant & b\left|\frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}_1}{||\boldsymbol{\beta}_1||_2^2} - \frac{\boldsymbol{\beta}_0^T\boldsymbol{\beta}_2}{||\boldsymbol{\beta}_2||_2^2}\right| + c_1\mathrm{E}|w| * \frac{|\boldsymbol{\beta}_0^T\boldsymbol{\beta}_2|}{||\boldsymbol{\beta}_2||_2^2} * |\sigma_{\boldsymbol{\beta}_1} - \sigma_{\boldsymbol{\beta}_2}| + c_1\frac{|\boldsymbol{\beta}_0^T\boldsymbol{\beta}_2|}{||\boldsymbol{\beta}_2||_2^2} * \left|\frac{(\boldsymbol{\beta}_0^T\boldsymbol{\beta}_1)}{||\boldsymbol{\beta}_1||_2^2}t - \frac{(\boldsymbol{\beta}_0^T\boldsymbol{\beta}_2)}{||\boldsymbol{\beta}_2||_2^2}t\right| \\
\triangleq & |E_1| + |E_2| + |E_3|,
\end{aligned}
$$

where the definition of E_k, $k = 1, \cdots, 3$, is clear from the context. As discussed above, we have that $|E_1| \leqslant 3b||\boldsymbol{\beta}_0||_2 * ||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2||_2$, $|E_2| \leqslant c_2 * ||\boldsymbol{\beta}_0||\sqrt{||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2||_2}$, and $|E_3| \leqslant c_2|t| * ||\boldsymbol{\beta}_0||^2 * ||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2||_2$ for some constant $c_2 > 0$. In conclusion, for some positive constant C, we have $\left|G^{(1)}(t|\boldsymbol{\beta}_1) - G^{(1)}(t|\boldsymbol{\beta}_2)\right| \leqslant C\Big[|t| * ||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2||_2 + \sqrt{||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2||_2}\Big]$. Then applying similar techniques as those in Section S9.4, we can validate Assumption (A5)-(c).

S10 Algorithms and additional numerical results

S10.1 Pseudo Codes for the algorithms in Section 4.1

In this subsection, we provide pseudo codes for the algorithms introduced in Section 4.1. Algorithm 1 is the main algorithm for solving the penalized high-dimensional profiled estimating equation for the initial estimator $\widehat{\boldsymbol{\beta}}$. It extends the proximal algorithm [Nesterov, 2007, Agarwal et al., 2012] to estimate the profiled semiparametric estimator. Algorithm 2 describes the details of the projection step in Algorithm 1, using an algorithm introduced in Duchi et al. [2008].

Algorithm 1 An algorithm for solving the penalized profiled estimating equation.
Input: initial value $\boldsymbol{\beta}^0$, λ, γ_u, data $\{\boldsymbol{x}_i, \widetilde{Y}_i\}_{i=1}^n$

Set $t = 1$, $\boldsymbol{\beta}^t = \boldsymbol{\beta}^{t-1} = \boldsymbol{\beta}^0$, coef.err $= ||\boldsymbol{\beta}^0||_2 + 1$, model.err^{t-1} =model.err^{t-2} = Var$(\widetilde{Y}_i)$.
while coef.err $> 0.01 * ||\boldsymbol{\beta}^{t-1}||_2$ or model.err^{t-1} < model.err^{t-2} **do**
    $h^t \leftarrow 0.9n^{-1/6} \min\{\mathrm{std}(\boldsymbol{x}_i^T\boldsymbol{\beta}^t), \mathrm{IQR}(\boldsymbol{x}_i^T\boldsymbol{\beta}^t)/1.34\}$.
    $w_{ij}^t \leftarrow K\left(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}^t - \boldsymbol{x}_j^T\boldsymbol{\beta}^t}{h^t}\right)$; $w_{ij}'^t \leftarrow (h^t)^{-1}K'\left(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}^t - \boldsymbol{x}_j^T\boldsymbol{\beta}^t}{h^t}\right)$.
    $\widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}^t|\boldsymbol{\beta}^t) \leftarrow \frac{\sum_{j\neq i} w_{ij}^t \widetilde{Y}_j}{\sum_{j\neq i} w_{ij}^t}$.
    $\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}^t|\boldsymbol{\beta}^t) \leftarrow \frac{\sum_{j\neq i} w_{ij}'^t \widetilde{Y}_j}{\sum_{j\neq i} w_{ij}^t} - \widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}^t|\boldsymbol{\beta}^t) * \sum_{j\neq i} w_{ij}'^t$.
    $\widehat{\mathrm{E}}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}^t) \leftarrow \frac{\sum_{j\neq i} w_{ij}^t \boldsymbol{x}_j}{\sum_{j\neq i} w_{ij}^t}$; $\widehat{\boldsymbol{x}}_i^t \leftarrow \boldsymbol{x}_i - \widehat{\mathrm{E}}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}^t)$.
    model.err$^t \leftarrow \frac{1}{n}\sum_{i=1}^n[\widetilde{Y}_i - \widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}^t|\boldsymbol{\beta}^t)]^2$.
    $\boldsymbol{\beta}_{-1}^{t+1} \leftarrow \underset{\boldsymbol{\beta}_{-1}\in\mathbb{R}^{p-1}:||\boldsymbol{\beta}_{-1}||_1\leqslant\rho}{\arg\min} \frac{\gamma_u}{2}||\boldsymbol{\beta}_{-1} - \boldsymbol{\beta}_{-1}^t||_2^2 + [\boldsymbol{S}_n(\boldsymbol{\beta}^t, \widehat{G}, \widehat{\mathrm{E}})]^T(\boldsymbol{\beta}_{-1} - \boldsymbol{\beta}_{-1}^t) + \lambda||\boldsymbol{\beta}_{-1}||_1$, by (21) and Algorithm 2.
    $\boldsymbol{\beta}^{t+1} \leftarrow (1, (\boldsymbol{\beta}_{-1}^{t+1})^T)^T$.
    coef.err$\leftarrow ||\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t||_2$.
    $t \leftarrow t+1$, $\gamma_u \leftarrow 2 * \gamma_u$.
end while
Output $\boldsymbol{\beta}^t$.

Algorithm 2 An algorithm for projecting $\boldsymbol{\beta}$ onto the L_1-ball: $\{\boldsymbol{\beta} : ||\boldsymbol{\beta}||_1 \leqslant \rho\}$.
Input: initial value $\boldsymbol{\beta}$, ρ

if $||\boldsymbol{\beta}||_1 \leqslant \rho$ **then**
    Output $\boldsymbol{\beta}$.
else
    Sort $\{|\beta_j|\}_{j=1}^p$ into $b_1 \geqslant b_2 \geqslant \cdots \geqslant b_p$.
    Find $J = \max\{2 \leqslant j \leqslant p : b_j - \frac{(\sum_{r=1}^j b_r) - \rho}{j} > 0\}$, and $\delta = \frac{1}{J}[(\sum_{r=1}^J b_r) - \rho]$.
    Output $\boldsymbol{\beta}^o = T_s(\boldsymbol{\beta}, \delta)$.
end if

S10.2 Computation of $\mathbf{d_j}(\boldsymbol{\beta}, \eta)$

In Section 2.3, we introduce a nodewise Dantzig estimator $\boldsymbol{d}_j(\widehat{\boldsymbol{\beta}}, \eta)$, as defined in (11), to obtain the approximate inverse of $\nabla\boldsymbol{S}_n(\widehat{\boldsymbol{\beta}}, \widehat{G}, \widehat{\mathrm{E}})$. This estimator can be solved via a linear programming problem as follows:

$$\min_{\boldsymbol{\xi}^+,\boldsymbol{\xi}^- \in \mathbb{R}^{p-2}} ||\boldsymbol{\xi}^+||_1 + ||\boldsymbol{\xi}^-||_1 \text{ subject to } \boldsymbol{\xi}^+ \geqslant 0, \boldsymbol{\xi}^- \geqslant 0, \text{ and} \tag{S54}$$

$$\frac{1}{n}\sum_{i=1}^{n}[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2\widehat{x}_{i,k}\widehat{\boldsymbol{x}}_{i,-j*}^T(\boldsymbol{\xi}^+ - \boldsymbol{\xi}^-) \geqslant \frac{1}{n}\sum_{i=1}^{n}[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2\widehat{x}_{i,j}\widehat{x}_{i,k} - \eta, \text{ for all } k \neq 1, k \neq j,$$

$$\frac{1}{n}\sum_{i=1}^{n}[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2\widehat{x}_{i,k}\widehat{\boldsymbol{x}}_{i,-j*}^T(\boldsymbol{\xi}^+ - \boldsymbol{\xi}^-) \leqslant \frac{1}{n}\sum_{i=1}^{n}[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})]^2\widehat{x}_{i,j}\widehat{x}_{i,k} + \eta, \text{ for all } k \neq 1, k \neq j,$$

for any given $j \in \{2, \cdots, p\}$. Then $(\boldsymbol{\xi}^+ - \boldsymbol{\xi}^-)$ is an estimator of $\boldsymbol{d}_j$. In our numerical analysis, we apply the function "lp" in the R package `lpSolve` [Berkelaar and others, 2015] for linear programming.

S10.3 Additional numerical results

Example 1 (Effect of tuning parameter η). We compare inference performance with different choices of η ($\eta = 15h$, $20h$, $25h$, and $30h$) and the λ selected by 5-fold cross-validation. We consider the same model as in Section 4.2 in the main paper, with $n = 300$ and $p = 200$.

Table S1 summarizes the average Type I errors and powers. We observe that inference performance is not very sensitive to η. Also, the choice $\eta = 25h$ leads to performance similar as that obtained using η chosen by cross-validation.

Table S1: Performance of the bootstrap procedure for simultaneous testing with different choices of η.

η	Type I error	Power				
	$\mathcal{G}_1$	$\mathcal{G}_2$	$\mathcal{G}_3$	$\mathcal{G}_4$	$\mathcal{G}_5$	$\mathcal{G}_6$
15h	3.6%	79.6%	93.2%	94.4%	94.8%	100%
20h	5.0%	92.8%	95.6%	97.0%	97.6%	100%
25h	5.6%	96.4%	96.2%	97.8%	98.6%	100%
30h	4.2%	96.8%	96.6%	98.2%	99.2%	100%
CV	4.6%	97.2%	96.8%	98.4%	99.0%	100%

Example 2 (Comparison with alternative algorithms). We compare the proposed semiparametric procedure with the nonparametric O-learning procedure [Zhao et al., 2012], and the decision list based approach [Zhang et al., 2018]. We use the "DTRlearn2" R package with the Gaussian kernel for O-learning [Chen et al., 2019] and the "listdtr" R package for the decision list approach [Zhang, 2016]. As the alternative procedures do not perform inference, our comparison is focused on estimating the optimal value function. Given an estimated decision rule indexed by $\widehat{\boldsymbol{\beta}}$, we can estimate the optimal value function by $\widehat{V}(\widehat{\boldsymbol{\beta}}) = \frac{\sum_{i=1}^n I[A_i = d(\boldsymbol{x}_i)]Y_i}{\sum_{i=1}^n I[A_i = d(\boldsymbol{x}_i)]}$.

We consider two different settings. The first setting (setting 1) corresponds to the index model in Section 4.2 in the main paper, for which the optimal value is 3.423 based an independent Monte Carlo simulation with 10^7 replicates. In the second setting (setting 2), $Y = 1 + \boldsymbol{x}^T\boldsymbol{\eta} + (A - \frac{1}{2})f_0(\boldsymbol{x}) + \epsilon$, where $\epsilon \sim N(0,1)$, $A \sim$ Bernoulli(0.5), $\boldsymbol{x} = (x_1, \cdots, x_p)^T$ has elements independently distributed as Uniform$(-1,1)$, $\boldsymbol{\eta} = (2, 1, 0.5, 0, \cdots, 0)^T$ and $f_0(\boldsymbol{x}) = 20(1 - x_1^2 - x_2^2)(x_1^2 + x_2^2 - 0.36)$. The optimal value of setting 2 is 2.443, based on an independent Monte Carlo simulation with 10^7 replicates.

Table S2 summarizes the average bias and standard error for estimating the optimal values for the two settings for $p = 200, 800$ and $n = 300, 500$. It also reports the average match ratio (MR). MR is estimated as the percentage of times the estimated optimal decision rule coincides with the true optimal decision rule, the latter of which is computed using an independent sample of size 10^4. Due to the computational cost, for the decision list based estimators, we run 200 simulations. For the other two estimators, the results are based on 500 simulation runs.

We have the following observations. (1) In setting 1, our proposed method has smaller biases for estimating the optimal value comparing with the two other approaches. This is likely due to the fact the proposed method is semiparametric. In contrast, the other two approaches do not make use of the model structure in estimating the optimal decision rule. (2) In setting 2, O-learning has smaller bias for estimating the optimal value. It is noted that

Table S2: Estimated bias (with standard error in the parentheses) for the optimal value and the average match ratios

n	p		New	O-learning	List learning
Setting 1					
300	200	value	-0.034 (0.008)	0.195 (0.011)	-0.258 (0.010)
		MR	93.76%	75.89%	79.91%
	800	value	-0.055 (0.008)	0.277 (0.010)	-0.300 (0.028)
		MR	92.67%	63.70%	79.05%
500	200	value	-0.035 (0.006)	0.172 (0.009)	-0.289 (0.017)
		MR	95.51%	80.82%	81.43%
	800	value	-0.032 (0.006)	0.288 (0.008)	-0.272 (0.017)
		MR	94.75%	67.93%	80.77%
Setting 2					
300	200	value	-0.642 (0.010)	0.269 (0.010)	-0.676 (0.012)
		MR	50.60%	49.98%	49.98%
	800	value	-0.649 (0.009)	0.464 (0.009)	-0.647 (0.024)
		MR	50.78%	49.98%	50.07%
500	200	value	-0.673 (0.007)	0.141 (0.010)	-0.689 (0.020)
		MR	50.52%	49.98%	49.95%
	800	value	-0.655 (0.008)	0.434 (0.007)	-0.684 (0.017)
		MR	50.68%	49.96%	50.07%

in this setting the model does not have the index form and hence the proposed semiparametric procedure is based on a misspecified model. (3) In both settings, the performance of O-learning deteriorates as p gets larger while the performance of the new method is stable.

Example 3 (Correlated design with discrete covariates). In this example, the covariates include three discrete components, which are independent and uniformly distributed on the set $\{-1, 0, 1\}$. All the other covariates follow a $(p-3)$-dimensional multivariate normal distribution with mean zero and covariance matrix $\mathbf{\Sigma}$, with $\Sigma_{i,j} = 0.5^{|i-j|}$. The three discrete variables are the fifth and the last two of the p covariates. The model has the same form as the example in Section 4.2 of the main paper and has $\boldsymbol{\beta}_0 = (1, -1, -0.8, 0.6, -0.5, 0, \cdots, 0)$.

Table S3 summarizes the estimation results for $n = 300, 500$ and $p = 200, 800$, based on 500 simulations. We observe that the proposed profiled estimator has satisfactory performance in this experiment.

Table S3: Performance of the estimator for the correlated design with discrete covariates

n	p	l_1 error	l_2 error	False Negative	False Positive
300	200	0.81 (0.02)	0.32 (0.00)	0.00 (0.00)	8.31 (0.30)
	800	1.10 (0.02)	0.41 (0.01)	0.01 (0.01)	14.82 (0.61)
500	200	0.53 (0.01)	0.22 (0.00)	0.00 (0.00)	6.59 (0.25)
	800	0.73 (0.01)	0.28 (0.00)	0.00 (0.00)	12.48 (0.46)

Next we investigate the proposed wild bootstrap inference procedure for testing group hypotheses with $\mathcal{G}_1 = \{6, 7, 8, 9\}$, $\mathcal{G}_2 = \{5, 6, 7, 8, 9\}$, $\mathcal{G}_3 = \{4, 6, 7, 8, 9\}$, $\mathcal{G}_4 = \{4, 5, 6, 7, 8, 9\}$, $\mathcal{G}_5 = \{3, 6, 7, 8, 9\}$ and $\mathcal{G}_6 = \{2, 6, 7, 8, 9\}$. Note that $\mathcal{G}_2$ includes a discrete variable. Table S4 summarizes the results based on 1000 bootstrap samples and 500 simulation runs. We observe that the estimated type I errors and powers are reasonable for all scenarios.

Table S4: Performance of the wild bootstrap inference procedure for the correlated design with discrete covariates.

n	p	Type I error	Power				
		$\mathcal{G}_1$	$\mathcal{G}_2$	$\mathcal{G}_3$	$\mathcal{G}_4$	$\mathcal{G}_5$	$\mathcal{G}_6$
300	200	5.6%	98.8%	96.8%	97.2%	100%	100%
	800	6.6%	84.4%	86.6%	88.0%	99.8%	91.8%
500	200	5.0%	96.6%	97.8%	98.0%	100%	100%
	800	7.2%	89.8%	94.8%	95.6%	100%	92.2%

Example 4 (Addition results for the real-data example in Section 5 of the main paper). In Table S5, we report the estimated coefficients for the variables in Table 3. In the table, "insulin" stands for fasting insulin, "Cr" stands for creatinine, and "waist" stands for waist circumference.

Table S5: Real data analysis: profiled estimator for variables in Table 3 of the main paper

Variable	fasting insulin	creatinine	BMI	waist
Coef	1	0.0011	-0.0070	-0.0047
Variable	HbA_{1c}	HomaS	Cr:insulin	Cr:BMI
Coef	-0.0519	0.0071	0	0
Variable	Cr:waist	Cr:HbA_{1c}	Cr:HomaS	insulin:BMI
Coef	0	-0.0171	0.0110	0
Variable	insulin:waist	insulin:HbA_{1c}	insulin:HomaS	BMI:waist
Coef	0	0	0	-0.0046
Variable	BMI:HbA_{1c}	BMI:HomaS	waist:HbA_{1c}	waist:HomaS
Coef	-0.0349	0.0051	-0.0399	0.0067
Variable	HbA_{1c}:HomaS	LDL-C	total cholesterol	age
Coef	-0.0002	0.0030	0.0032	0.0022
Variable	weight			
Coef	0			

As an example of using the estimated model to interpret the covariate effect on the outcome. we consider the effect of baseline HbA_{1c} on the outcome of receiving the recommended treatment. Figure 1 plots the $\widehat{G}(\boldsymbol{x}^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})$ versus baseline HbA_{1c} while fixing all the other covariates at their respective sample averages. The plot suggests that for such an average patient, receiving pioglitazone (treatment 0) is likely to reduce the level of HbA_{1c}, and larger benefit is expected for patients with a smaller value of baseline HbA_{1c}.

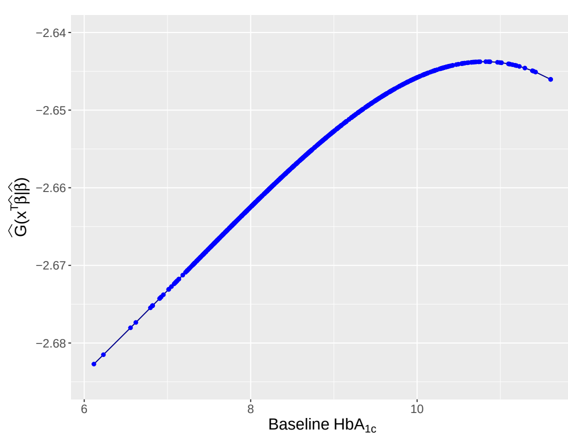

Figure 1: Plot of $\widehat{G}(\boldsymbol{x}^T\widehat{\boldsymbol{\beta}}|\widehat{\boldsymbol{\beta}})$ versus baseline HbA_{1c} while fixing all the other covariates at their respective sample averages.

Example 5 (Extension to observational studies). We investigate the procedure proposed in Section 6 of the main paper for observation studies. We consider the same model as in Section 4.2 of the main paper, except that A is generated according to $P(A = 1|\boldsymbol{x}) = [1 + \exp(-\boldsymbol{x}^T\boldsymbol{\xi})]^{-1}$, where $\boldsymbol{\xi} = (0.2, 0.2, -0.4, 0, \cdots, 0)^T$. We estimate the propensity score via L_1-regularized logistic regression. Table S6 suggests the promising performance for the proposed estimator for observational studies.

Table S6: Performance of the penalized profile least-squares estimator

n	p	l_1 error	l_2 error	False Negative	False Positive
300	200	1.00 (0.01)	0.42 (0.00)	0.06 (0.01)	9.13 (0.30)
	800	1.26 (0.02)	0.49 (0.00)	0.10 (0.01)	16.29 (0.62)
500	200	0.72 (0.01)	0.30 (0.00)	0.00 (0.00)	8.32 (0.26)
	800	0.97 (0.01)	0.39 (0.00)	0.01 (0.00)	15.18 (0.48)

References

R. Adamczak and P. Wolff. Concentration inequalities for non-lipschitz functions with bounded derivatives of higher order. *Probability Theory and Related Fields*, 162(3):531–586, 2015.

A. Agarwal, S. Negahban, and M. J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics*, 40(5):2452–2482, 2012.

A. Belloni, V. Chernozhukov, and K. Kato. Uniform post-selection inference for least absolute deviation regression and other z-estimation problems. *Biometrika*, 102(1):77–94, 2015.

M. Berkelaar and others. *lpSolve: Interface to 'Lp_solve' v. 5.5 to Solve Linear/Integer Programs*, 2015. R package version 5.6.13.

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer Series in Statistics. Springer Berlin Heidelberg, 2011.

B. Chakraborty, E. B. Laber, and Y.-Q. Zhao. Inference about the expected performance of a data-driven dynamic treatment regime. *Clinical Trials*, 11(4):408–417, 2014.

Y. Chen, Y. Liu, D. Zeng, and Y. Wang. *DTRlearn2: Statistical Learning Methods for Optimizing Dynamic Treatment Regimes*, 2019. R package version 1.0.

V. Chernozhukov, D. Chetverikov, and K. Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819, 2013.

J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine Learning*, pages 272–279, 2008.

J. Horowitz. *Semiparametric Methods in Econometrics*. Lecture Notes in Statistics. Springer New York, 2012.

H. Ichimura. Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58(1-2):71–120, 1993.

P.-L. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664, 2012.

Y. Nesterov. Gradient methods for minimizing composite objective function. Core discussion papers, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2007.

S. Van de Geer, P. Bühlmann, Y. Ritov, R. Dezeure, et al. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3): 1166–1202, 2014.

A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, 1996.

M. J. Wainwright. Basic tail and concentration bounds, 2015.

J.-L. Wang, L. Xue, L. Zhu, Y. S. Chong, et al. Estimation for a partial-linear single-index model. *The Annals of Statistics*, 38(1):246–274, 2010.

L. A. Wasserman. Stein's method and the bootstrap in low and high dimensions : A tutorial. 2014.

Y. Yu and D. Ruppert. Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, 97(460):1042–1054, 2002.

Y. Zhang. *listdtr: List-Based Rules for Dynamic Treatment Regimes*, 2016. R package version 1.0.

Y. Zhang, E. B. Laber, M. Davidian, and A. A. Tsiatis. Interpretable dynamic treatment regimes. *Journal of the American Statistical Association*, 113(524):1541–1549, 2018.

Y. Zhao, D. Zeng, A. J. Rush, and M. R. Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.

L. Zhu and L. Xue. Empirical likelihood confidence regions in a partially linear single-index model. *Journal of the Royal Statistical Society: Series B*, 68(3):549–570, 2006.